

NOT YET SCHEDULED FOR ORAL ARGUMENT**Appeal No. 17-7035****(Consolidated with Appeal No. 17-7039)**

United States Court of Appeals**FOR THE DISTRICT OF COLUMBIA CIRCUIT**

American Society for Testing and Materials; National Fire Protection Association, Inc.; and American Society of Heating, Refrigerating, and Air Conditioning Engineers, Inc.,

Appellees,

v.

Public.Resource.Org, Inc.,

Appellant.

Appeal from the United States District Court for the District of Columbia

Hon. Tanya S. Chutkan

1:13-cv-1215-TSC

1:14-cv-0857-TSC

**PUBLIC APPENDIX – MATERIAL UNDER SEAL
IN SEPARATE SUPPLEMENT
VOLUME 5 (JA2333-JA2881)**

Andrew P. Bridges
abridges@fenwick.com

Matthew B. Becker
mbecker@fenwick.com

Fenwick & West LLP
555 California Street
San Francisco, CA 94104
Phone: (415) 875-2300

Corynne McSherry
corynne@eff.org

Mitchell L. Stoltz
mitch@eff.org

Electronic Frontier Fndn.
815 Eddy Street
San Francisco, CA 94109
Phone: (415) 436-9333

David Halperin
davidhalperindc@gmail.com

1530 P Street NW
Washington, DC 20005
Phone: (202) 905-3434

Attorneys for Appellant Public.Resource.Org, Inc.
Additional Counsel Listed on Inside Cover

Michael J. Songer
John I. Stewart Jr.
Clifton S. Elgarten
Mark Thomson
CROWELL & MORING LLP
1001 Pennsylvania Avenue, NW
Washington, DC 20004-2505
Phone: (202) 624-2500
celgarten@crowell.com

Attorneys for Plaintiffs-Appellees
American Educational Research
Association, Inc.; American Psychological
Association, Inc.; and National Council
On Measurement In Education, Inc.

Allyson N. Ho
Morgan, Lewis & Bockius LLP
1717 Main Street, Suite 3200
Dallas, TX 75201
Phone: (214) 466-4180
allyson.ho@morganlewis.com

J. Kevin Fee
Jordana S. Rubel
Morgan, Lewis & Bockius LLP
1111 Pennsylvania Avenue, N.W.
Washington, D.C. 20004
Phone: (202) 739-5353
kevin.fee@morganlewis.com
jordana.rubel@morganlewis.com

Attorneys for American Society for
Testing and Materials d/b/a/ ASTM
International

Donald B. Verrilli, Jr.
Munger, Tolles & Olson LLP
1155 F. Street, N.W., 7th Floor
Washington, D.C. 20004
donald.verrilli@mto.com

Kelly M. Klaus
Rose Leda Ehler
Munger, Tolles & Olson LLP
560 Mission Street, 27th Floor
San Francisco, CA 94105
Phone: (415) 512-4000
kelly.klaus@mto.com
rose.ehler@mto.com

Attorneys for National Fire Protection
Association, Inc.

Anne Voigts
Joseph R. Wetzel
King & Spalding LLP
101 Second Street, Suite 2300
San Francisco, CA 94105
Phone: (415) 318-1211
avoigts@kslaw.com
jwetzel@kslaw.com

J. Blake Cunningham
King & Spalding LLP
500 West 2nd Street, Suite 1800
Austin, TX 78701
Phone: (512) 457-2000
bcunningham@kslaw.com

Attorneys for American Society of
Heating, Refrigerating, and Air
Conditioning Engineers, Inc.

TABLE OF CONTENTS

Page

AMERICAN SOCIETY FOR TESTING AND MATERIALS ET AL. V.
PUBLIC.RESOURCE.ORG, INC.

VOLUME 1:

U.S. District Court for the District of Columbia Docket Sheet, AMERICAN SOCIETY FOR TESTING AND MATERIALS et al. v. PUBLIC.RESOURCE.ORG, INC. ASTM-DKT-000 DOCKET	JA1
Complaint, AMERICAN SOCIETY FOR TESTING AND MATERIALS et al. v. PUBLIC.RESOURCE.ORG ASTM-DKT-001	JA47
Exh A; ASTM COPYRIGHT REGISTRATIONS ASTM-DKT-001-1	JA57
Exh B; NATIONAL FIRE PROTECTION ASSOCIATION, INC. COPYRIGHT REGISTRATIONS ASTM-DKT-001-2	JA59
Exh C; AMERICAN SOCIETY OF HEATING, REFRIGERATION, REFRIGERATING AND AIR-CONDITIONING ENGINEERS, INC. COPYRIGHT REGISTRATIONS ASTM-DKT-001-3	JA61
Exh G; Incorporation by Reference: ASTM D4239: Standard Test Methods for Sulfur in the Analysis Sample of Coal and Coke Using High Temperature Tube Furnace Combustion Methods ASTM-DKT-001-7	JA65
PLAINTIFF NATIONAL FIRE PROTECTION ASSOCIATION, INC.'S OPPOSITION TO MOTION TO COMPEL DISCOVERY ASTM-DKT-046	JA121

TABLE OF CONTENTS
(Continued)

	Page
DECLARATION OF CHRISTIAN DUBAY IN SUPPORT OF PLAINTIFF NATIONAL FIRE PROTECTION ASSOCIATION ASTM-DKT-046-1	JA116
PLAINTIFFS' MEMORANDUM OF LAW IN SUPPORT OF THEIR MOTION FOR SUMMARY JUDGMENT AND FOR A PERMANENT INJUNCTION ASTM-DKT-118-01	JA138
DECLARATION OF DENNIS J. BERRY IN SUPPORT OF PLAINTIFFS' MOTION FOR SUMMARY JUDGMENT ASTM-DKT-118-03	JA144
Exh A; Certificate of Registration - National Electrical Code, 2011 Edition ASTM-DKT-118-03, Ex. A.....	JA148
Exh B; Certificate of Registration - National Electrical Code, 2014 Edition ASTM-DKT-118-03, Ex. B.....	JA150
Exh J; Email from Vacay Promo Team to Dennis Berry re: Ebay Violation?, dated 01/22/15 ASTM-DKT-118-03, Ex. J	JA152
Exh K; Email from Scott Schwartz to Dennis Berry re: Use of Electronic NEC, dated 10/13/15 ASTM-DKT-118-03, Ex. K.....	JA156
Declaration of Steven Cramer ASTM-DKT-118-04	JA158
Declaration of James Golinveaux ASTM-DKT-118-05	JA163
Declaration of Randy Jennings ASTM-DKT-118-06.....	JA167
Declaration of Thomas B. O'Brien, Jr. ASTM-DKT-118-07	JA172

TABLE OF CONTENTS
(Continued)

	Page
Exh 1; Certificate of Registration: ASTM D86-07 Standards Test Methods for Distillation of Petroleum Products at Atmospheric Pressure ASTM-DKT-118-07, Ex. 01	JA183
Exh 2; Certificate of Registration: ASTM D975-07 Standard Specification for Diesel Fuel Oils ASTM-DKT-118-07, Ex. 02	JA186
Exh 3; Alphanumeric List: ASTM Standards ASTM-DKT-118-07, Ex. 03	JA189
Exh 4; Certificate of Registration: 1999 Annual Book of ASTM Standards ASTM-DKT-118-07, Ex. 04	JA193
Exh 5; Form and Style for ASTM Standards ASTM-DKT-118-07, Ex. 05	JA196
Exh 6; Designation ASTM D 86-07; Standard Test Methods for Distillation of Petroleum Products at Atmospheric Pressure ASTM-DKT-118-07, Ex. 06	JA277
Exh 7; Designation ASTM D 975-07 Standard Specification for Diesel Fuel Oils ASTM-DKT-118-07, Ex. 07	JA309
Exh 8; Designation ASTM D 396-98 Standard Specification for Fuel Oils ASTM-DKT-118-07, Ex. 08	JA327
Exh 9; Standard Test Method for Density and Relative Density (Specific Gravity) of Liquids by Bingham Pycnometer ASTM-DKT-118-07, Ex. 09	JA333
Exh 17; ASTM D8607 Viewer ASTM-DKT-118-07, Ex. 17	JA339
Exh 18; Standard Consumer Safety Specification for Infant Walkers ASTM-DKT-118-07, Ex. 18	JA344

TABLE OF CONTENTS
(Continued)

	Page
Declaration of James T. Pauley in Support of Plaintiff's Motion for Summary Judgment	
ASTM-DKT-118-08	JA364
Declaration of Stephanie Reiniche	
ASTM-DKT-118-10	JA379
Exh 3; Certificate of Registration - ANSI/ASHRAE/IESNA Standard 90 1-2004, Energy Standard for Buildings Except Low-Rise Residential Buildings	
ASTM-DKT-118-10, EXH 3	JA387
Exh 4; Certificate of Registration - ANSI/ASHRAE/IESNA Standard 90 1-2007, Energy Standard for Buildings Except Low-Rise Residential Buildings	
ASTM-DKT-118-10, EXH 4	JA390
Exh 5; Certificate of Registration - ANSI/ASHRAE/IESNA Standard 90.1-2010, Energy Standard for Buildings Except Low-Rise Residential Buildings	
ASTM-DKT-118-10, EXH 5	JA393
Declaration of James Thomas	
ASTM-DKT-118-11	JA396
Declaration of Jordana S. Rubel	
ASTM-DKT-118-12	JA403
Exh 1; Expert Report of John C. Jarosz; dated 06/05/15	
ASTM-DKT-118-12, Ex. 01 (Material Under Seal)	JA409
Exh 2; Rule 30(b)(6) Deposition of Public.Resource.Org, dated 02/26/15	
ASTM-DKT-118-12, Ex. 02	JA524
Exh 3; Deposition of Carl Malamud, dated 02/27/15	
ASTM-DKT-118-12, Ex. 03 (Material Under Seal)	JA602
Exh 4; Rule 30(b)(6) Deposition of Rebecca Malamud, Public.Resource.Org, dated 11/13/14 (excerpts)	
ASTM-DKT-118-12, Ex. 04	JA706

TABLE OF CONTENTS
(Continued)

	Page
Exh 6; Rule 30(b)(6) Deposition of National Fire Protection Association, Inc., dated 04/01/15 (excerpts) ASTM-DKT-118-12, Ex. 06	JA715
Exh 7; Rule 30(b)(6) Deposition of ASTM Designee Stephanie Reiniche, dated 03/30/15 (excerpts) ASTM-DKT-118-12, Ex. 07	JA742
Exh 10; Correspondence from U.S. Dept. of the Interior to Carl Malamud, dated 09/08/15 ASTM-DKT-118-12, Ex. 10	JA757
Exh 12; dharlanuctcom 73 uploads ASTM-DKT-118-12, Ex. 12	JA766

VOLUME II:

Exh 16; Incorporation by Reference - ASTM Designation D 86-07 Standard Test Method for Distillation of Petroleum Products at Atmospheric Pressure ASTM-DKT-118-12, Ex. 16	JA771
Exh 19; Email from Carl Malamud to Rebecca Malamud re: SVG and MathML (India and NFPA / Q4); dated 01/04/14 ASTM-DKT-118-12, Ex. 19	JA801
Exh 23; NFPA NEC (20110 National Electrical Code (January 1, 2011) ASTM-DKT-118-12, Ex. 23	JA805
Exh 25; ASHRAE Standard - Energy Standard for Buildings Except Low-Rise Residential Buildings ASTM-DKT-118-13, Ex. 25	JA810
Exh 26; Incorporation by Reference - NFPA 70, NEC 2011 ASTM-DKT-118-13, Ex. 26	JA814

TABLE OF CONTENTS
(Continued)

	Page
Exh 27; Public Safety Standards United States (Federal Government) ASTM-DKT-118-13, Ex. 27	JA818
Exh 28; Public Safety Codes Incorporated by Law ASTM-DKT-118-13, Ex. 28	JA895
Exh 29; ASTM Designation D 86-07 Standard Test Method for Distillation of Petroleum Products at Atmospheric Pressure ASTM-DKT-118-13, Ex. 29	JA906
Exh 30; Project Update #8: A Prayer for Our Democracy ASTM-DKT-118-13, Ex. 30	JA930
Exh 31; Project Update #6: Meet the Code People ASTM-DKT-118-13, Ex. 31	JA934
Exh 32: Twelve Tables of Codes ASTM-DKT-118-13, Ex. 32	JA942
Exh 33: Email from Carl Malamud to Josh Greenberg re: Federal Register/Code of Federal Regulations, dated 08/24/11 ASTM-DKT-118-13, Ex. 33	JA956
Exh 34: Email from Carl Malamud re: suit; dated 08/09/13 ASTM-DKT-118-13, Ex. 34	JA972
Exh 38; downloads identifier chart ASTM-DKT-118-14, Ex. 38	JA974
Exh 39: Search Results - NFPA AND collection: additional collections ASTM-DKT-118-14, Ex. 39	JA988
Exh 44; ASTM D975 Standard Specification for Diesel Fuel Oils (2007) ASTM-DKT-118-16, Ex. 44	JA996
Exh 4; Rule 30(b)(6) Deposition of Donald P. Bliss, dated 03/03/15 (excerpts) ASTM-DKT-120-06 (Material Under Seal).....	JA999

TABLE OF CONTENTS
(Continued)

	Page
Exh 11; Rule 30(b)(6) Deposition of National Fire Protection Association, dated 03/31/15 (excerpts) ASTM-DKT-120-07 (Material Under Seal).....	JA1019
Exh 22: Comments of ASTM International, dated 04/15/12 ASTM-DKT-120-09 (Material Under Seal).....	JA1032
Exh 74; Form for Comments on NFPA Report on Proposals ASTM-DKT-120-11 (Material Under Seal).....	JA1040
Exh 75; Form for Proposals on NFPA Technical Committee Documents ASTM-DKT-120-12 (Material Under Seal).....	JA1043
Exh 76; Form for Proposals on NFPA Technical Committee Documents ASTM-DKT-120-13 (Material Under Seal).....	JA1045
Exh 80; ASTM 2011 Membership Renewal Invoice ASTM-DKT-120-14 (Material Under Seal).....	JA1049
Exh 83; 2010 ASTM International Committee Membership Application ASTM-DKT-120-16 (Material Under Seal).....	JA1052
Exh 140; Correspondence from Jeff, to Jim; re: 2012 Accomplishments and 2013 Objectives ASTM-DKT-120-30 (Material Under Seal).....	JA1054
Exh 141; Email from John Pace to Jeff Groves re Standards Summaries, dated 07/09/12 ASTM-DKT-120-31 (Material Under Seal).....	JA1058
Exh 142; Email from John Pace to Jeff Groves re Standards Summaries, dated 07/10/12 ASTM-DKT-120-32 (Material Under Seal).....	JA1061
Exh 146; Email from James Thomas to Mary McKiel re: ANSI IPRPC: Malamud update 01/15 ASTM-DKT-120-33 (Material Under Seal).....	JA1065

TABLE OF CONTENTS
(Continued)

	Page
Memorandum of Points & Authorities in Support of Public.Resource.Org's Motion for Summary Judgment and Opposition to Plaintiff's Motion for Summary Judgment and Permanent Injunction ASTM-DKT-121-1	JA1068
Declaration of Carl Malamud in Support of Public.Resource.Org's Motion for Summary Judgment ASTM-DKT-121-5	JA1070
Exh 2; Public Safety Standards, United States (Federal Government) ASTM-DKT-122-1, EXH 002.....	JA1081
Exh 5; Rule 30(b)(6) Deposition of Steven Comstock, dated 03/05/15 (excerpts) ASTM-DKT-122-1, EXH 005.....	JA1136
Exh 6; Rule 30(b)(6) Deposition of National Fire Protection Association, dated 04/01/15 (excerpts) ASTM-DKT-122-1, EXH 006.....	JA1149
Exh 8; Rule 30(b)(6) Deposition of American Standards Society for Testing and Materials, dated 03/04/15 (excerpts) ASTM-DKT-122-1, EXH 008.....	JA1165
Exh 9; Deposition of John C. Jarosz, dated 08/27/15 (excerpts) ASTM-DKT-122-1, EXH 009.....	JA1184
Exh 12; Rule 30(b)(6) Deposition of American Society of Heating, Refrigerating, and Air-Conditioning Engineers, Inc., designee Stephanie Reiniche, dated 03/30/15 (excerpts) ASTM-DKT-122-2 , EXH 012.....	JA1210
Exh 13; Rule 30(b)(6) Deposition of American Society for Testing and Materials, designee Daniel Smith, dated 07/24/15 (excerpts) ASTM-DKT-122-2 , EXH 013.....	JA1242

TABLE OF CONTENTS
(Continued)

Page

VOLUME III:

Exh 14; Certificate of Registration, 1983 Book of ASTM Standards Section 3 Volume 03.01 Metals - Mechanical Testing; Elevated and Low Temperature Tests ASTM-DKT-122-2, EXH 014.....	JA1271
Exh 15; Certificate of Registration, NFPA 1 Uniform Fire Code 2003 Edition ASTM-DKT-122-2, EXH 015.....	JA1440
Exh 16; Certificate of Registration; 1993 ASHRAE Handbook -- Fundamentals, Inch-Pound Edition ASTM-DKT-122-2, EXH 016.....	JA1481
Exh 23; Presentation - ASTM Standards, Regulations and Trade ASTM-DKT-122-3, EXH 023.....	JA1490
Exh 24; Email from Sarah Petre to Jeff Grove, re: ASTM Follow Up on S1492, dated 10/04/12 ASTM-DKT-122-3, EXH 024.....	JA1513
Exh 26; Incorporation by Reference Public Workshop ASTM-DKT-122-3, EXH 026.....	JA1518
Exh 28; CM Submittal Form ASTM-DKT-122-3, EXH 028.....	JA1534
Exh 44; Application for Project Committee Organizational Representative Membership ASTM-DKT-122-4, EXH 044.....	JA1536
Exh 48; Application for Membership on ASHRAE Standard or Guideline Project Committee ASTM-DKT-122-4, EXH 048.....	JA1540
Exh 49; Memorandum of Understanding Between The United States Dept. of Energy and The American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc. ASTM-DKT-122-4, EXH 049.....	JA1543

TABLE OF CONTENTS
(Continued)

	Page
Exh 50: Marketing Task Group, TG Meeting Report, 06/26/04 ASTM-DKT-122-4, EXH 050.....	JA1547
Exh 51; Email from Steve Ferguson to Doug Read et al., re: IECC and 90.1, dated 12/17/09 ASTM-DKT-122-4, EXH 051.....	JA1563
Exh 52; ASHRAE Government Affairs: Technical Expertise to Policymakers ASTM-DKT-122-4, EXH 052.....	JA1566
Exh 73: Form for Comments on NFPA Report on Proposals 2000 November Association Technical Meeting ASTM-DKT-122-5, EXH 073.....	JA1599
Exh 77: Intellectual Property Policy of ASTM, Introduction ASTM-DKT-122-5, EXH 077.....	JA1602
Exh 78: Intellectual Property Policy of ASTM International ("Policy") ASTM-DKT-122-5, EXH 078.....	JA1611
Exh 79: Intellectual Property Policy of ASTM International ("Policy") ASTM-DKT-122-5, EXH 079.....	JA1616
Exh 95: How To: Standards Writing 101, New Standards ASTM-DKT-122-6, EXH 095.....	JA1623
Exh 96: Expert Report of James R. Fruchterman ASTM-DKT-122-6, EXH 096.....	JA1628
Exh 97; ASTM Editions Incorporated by Reference ASTM-DKT-122-6, EXH 097.....	JA1689
Exh 98; NFPA Editions Incorporated by Reference ASTM-DKT-122-6, EXH 098.....	JA1720
Exh 99; ASHRAE Editions Incorporated by Reference ASTM-DKT-122-6, EXH 099.....	JA1723

**TABLE OF CONTENTS
(Continued)**

	Page
Exh 104; Capitol Hill Event to Feature Policy and Business Leader Insights on Voluntary Standards and Conformance ASTM-DKT-122-7, EXH 104.....	JA1725
Exh 106: ASTM International, Public Policy & Corporate Outreach ASTM-DKT-122-7, EXH 106.....	JA1728
Exh 123; NFPA 70 National Electrical Code, 2011 Edition, Errata No. 70-11-1 ASTM-DKT-122-8, EXH 123.....	JA1754
Exh 124; NFPA 70 National Electrical Code, 2011 Edition, Errata No. 70-11-2 ASTM-DKT-122-8, EXH 124.....	JA1757
Exh 125; An Introduction to the NFPA Standards Development Process ASTM-DKT-122-8, EXH 125.....	JA1759

VOLUME IV:

Exh 126: ASHRAE Standards Committee; Procedures for ASHRAE Standards Actions PASA ASTM-DKT-122-8, EXH 126.....	JA1780
Exh 127; Form for Proposals for 2011 National Electrical Code ASTM-DKT-122-8, EXH 127.....	JA1820
Exh 128; Form for Proposals for 2008 National Electrical Code ASTM-DKT-122-8, EXH 128.....	JA1822
Exh 129; Form for Proposals for 2011 National Electrical Code ASTM-DKT-122-8, EXH 129 (Material Under Seal)	JA1824
Exh 130; Email from John Pace to Kathe Hooper re: Question related to copyright, dated 03/24/09 ASTM-DKT-122-8, EXH 130.....	JA1833

TABLE OF CONTENTS
(Continued)

	Page
Exh 131; Email from Kathe Hooper to Victor Palacios re: Request (nao), dated 07/09/09 ASTM-DKT-122-8, EXH 131.....	JA1838
Exh 132; ASTM International, Register My Account ASTM-DKT-122-8, EXH 132.....	JA1843
Exh 133; ASTM International, Checkout - Your Address ASTM-DKT-122-8, EXH 133.....	JA1845
Exh 134; ASTM International, Reading Room ASTM-DKT-122-8, EXH 134.....	JA1847
Exh 135; "The purpose of this site...." ASTM-DKT-122-8, EXH 135.....	JA1849
Exh 136; ASTM International, ASTM License Agreement ASTM-DKT-122-8, EXH 136.....	JA1852
Exh 137; "Please indicate your acceptance..." ASTM-DKT-122-8, EXH 137.....	JA1854
Exh 138; NFPA.Org/Login National Fire Protection Association web page ASTM-DKT-122-8, EXH 138.....	JA1857
Exh 139; ASHRAE Shaping Tomorrow's Built Environment Today ASTM-DKT-122-8, EXH 139.....	JA1859
Exh 143; ASTM License Agreement (Reading Room) ASTM-DKT-122-9, EXH 143.....	JA1861
Exh 154; NFPA Standards Development Site; Public Comment Stage ASTM-DKT-122-9, EXH 154.....	JA1864
Exh 155; National Archives, Incorporation by Reference ASTM-DKT-122-9, EXH 155.....	JA1877

TABLE OF CONTENTS
(Continued)

	Page
Exh 1; Deposition of John C. Jarosz, dated 08/27/15 (excerpts) ASTM-DKT-124-3	JA1882
Exh 3; Be confident your electrical work complies with California law (Gmail) ASTM-DKT-124-5	JA1930
Plaintiff's Opposition to Defendant's Motion for Summary Judgment and Reply Memorandum of Law in Support of their Motion for Summary Judgment and for a Permanent Injunction ASTM-DKT-155	JA1933
Declaration of Steve Comstock ASTM-DKT-155-5	JA1935
Declaration of Christian Dubay in Support of Plaintiffs' Motion for Summary Judgment ASTM-DKT-155-6	JA1940
Supplemental Declaration of Thomas B. O'Brien, Jr. ASTM-DKT-155-7	JA1945
Exh 1; Deposition of James Fruchterman, dated 07/31/15 (excerpts) ASTM-DKT-155-8, Ex. 01	JA1950
Exh 4; Rule 30(b)(6) Deposition of American Standards Society for Testing and Materials, by designee Jeffrey Grove, dated 03/04/15 (excerpts) ASTM-DKT-155-8, Ex. 04	JA1981
Exh 7; Rule 30(b)(6) Deposition of Steven Comstock, dated 03/05/15 (excerpts) ASTM-DKT-155-8, Ex. 07	JA1990
Supplemental Declaration of Carl Malamud in Further Support of Defendant's Motion for Summary Judgment ASTM-DKT-164-8	JA2005
Exh 1; Executive Office of the President, Office of Management and Budget, OMB Circular A-119 ASTM-DKT-169-1	JA2007

**TABLE OF CONTENTS
(Continued)**

	Page
Order re: Motion to Strike Expert Report of John C. Jarosz ASTM-DKT-172	JA2051
Transcript of Motions Hearing before the Honorable Tanya S. Chutkan, dated 09/12/16 ASTM-DKT-173	JA2054
Memorandum Opinion, Dated 02/02/17 ASTM-DKT-175	JA2059
Order ASTM-DKT-176	JA2114
Notice of Appeal by Defendant-Counterclaimant Public.Resource.Org, Inc., dated 02/15/17 ASTM-DKT-177	JA2115
Amended Order ASTM-DKT-182	JA2118
Amended Notice of Appeal by Defendant-Counterclaimant Public.Resource.Org, Inc. ASTM-DKT-183	JA2120

**AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, INC. ET AL.
V. PUBLIC.RESOURCE.ORG, INC.**

(VOLUME IV – CONTINUED)

U.S. District Court for the District of Columbia Docket Sheet, AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, INC. et al. v. PUBLIC.RESOURCE.ORG, INC. AERA-DKT-000 DOCKET	JA2123
---	--------

TABLE OF CONTENTS
(Continued)

	Page
Complaint, AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, INC. et al. v. PUBLIC.RESOURCE.ORG, INC. AERA-DKT-001.....	JA2158
Plaintiff's Reply and Affirmative Defenses AERA-DKT-014.....	JA2186
MPA in Support of Plaintiff's Motion for Summary Judgment AERA-DKT-060-01	JA2211
Plaintiff's Statement of Material Facts in Support of Motion for Summary Judgment AERA-DKT-060-02	JA2215
Exh T; Public.Resource.Org, Inc.'s Amended Responses to First Set of Interrogatories (Nos. 1-8) AERA-DKT-060-23	JA2217
Exh V-1; Standards for Educational and Psychological Testing AERA-DKT-060-25	JA2233

VOLUME V:

Exh V-2; 9. Testing Individuals of Diverse Linguistic Backgrounds AERA-DKT-060-26	JA2333
Exh Z; Deposition of James R. Fruchterman, 09/08/15, (excerpts) AERA-DKT-060-30	JA2436
Exh II; archive-ssh-80x24 screen capture image AERA-DKT-060-44	JA2458
Exh JJ; Email dated 12/16/13 from John S. Neikirk to Carl Malamud AERA-DKT-060-45	JA2460

TABLE OF CONTENTS
(Continued)

	Page
Exh KK; 12/19/13 Correspondence to John Neikirk from Carl Malamud AERA-DKT-060-46	JA2462
Exh MM: Memorandum dated 06/12/14 AERA-DKT-060-48	JA2465
Declaration of Marianne Ernesto in Support of Plaintiff's Motion for Summary Judgment AERA-DKT-060-49	JA2467
Exh VV; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 04/24/14 AERA-DKT-060-58	JA2477
Exh WW; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 09/10/14 AERA-DKT-060-59	JA2480
Exh XX; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 09/08/14 AERA-DKT-060-60	JA2483
Exh YY; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 04/21/14 AERA-DKT-060-61	JA2486
Exh ZZ; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 04/21/14 AERA-DKT-060-62	JA2488
Exh AAA; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 04/22/14 AERA-DKT-060-63	JA2491
Exh BBB; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 04/21/14 AERA-DKT-060-64	JA2496

TABLE OF CONTENTS
(Continued)

	Page
Exh CCC: Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 10/16/14 AERA-DKT-060-65	JA2499
Exh DDD: Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 04/28/14 AERA-DKT-060-66	JA2503
Exh EEE: Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 09/08/14 AERA-DKT-060-67	JA2506
Exh FFF: Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 10/21/14 AERA-DKT-060-68	JA2509
Exh GGG: Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 04/23/14 AERA-DKT-060-69	JA2512
Exh HHH: Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 04/24/14 AERA-DKT-060-70	JA2515
Declaration of Laress L. Wise in Support of Plaintiff's Motion for Summary Judgment AERA-DKT-060-73	JA2518
Declaration of Wayne Camara in Support of Plaintiff's Motion for Summary Judgment AERA-DKT-060-76	JA2544
Exh MMM; Email from John S. Neikirk to Wayne Camara re: Existing Standards, dated 02/14/14 AERA-DKT-060-77	JA2576

TABLE OF CONTENTS
(Continued)

	Page
Declaration of Felice J. Levine in Support of Plaintiff's Motion for Summary Judgment AERA-DKT-060-78	JA2579
Exh QQQ: Standards for Educational & Psychological Testing AERA-DKT-060-82	JA2588
Exh RRR; Copyright registration for "Standards for Educational and Psychological Testing" AERA-DKT-060-83	JA2590
Exh SSS; Certificate of Registration for "Standards for Educational and Psychological Testing" AERA-DKT-060-84	JA2593
Exh TTT-1; "Standards for Educational and Psychological Testing" AERA-DKT-060-85	JA2596
Declaration of Kurt F. Geisenger in Support of Plaintiff's Motion for Summary Judgment AERA-DKT-060-88	JA2696
Public Resource's Motion to Strike Declaration of Kurt F. Geisenger in Support of Plaintiff's Motion for Summary Judgment AERA-DKT-067.....	JA2744
Memorandum of Points and Authorities in Support of Public Resource's Motion to Strike Declaration of Kurt F. Geisenger in Support of Plaintiff's Motion for Summary Judgment AERA-DKT-067-01(Material Under Seal).....	JA2746
Declaration of Matthew Becker in Support of Public Resource's Motion to Strike Declaration of Kurt F. Geisenger in Support of Plaintiff's Motion for Summary Judgment AERA-DKT-067-02 (Material Under Seal).....	JA2769
Exh 1; Deposition of Kurt P. Geisinger, dated 09/10/15 (excerpts) AERA-DKT-067-03 (Material Under Seal).....	JA2774

TABLE OF CONTENTS
(Continued)

	Page
Exh 2; Deposition of Felice J. Levine, dated 05/04/15 (excerpts) AERA-DKT-067-04 (Material Under Seal).....	JA2799
Exh 3; Expert's Declaration and Report of Kurt F. Geisinger, dated 06/10/15 AERA-DKT-067-05	JA2804
Exh 4; Geisinger Expert Report - List of Materials Considered AERA-DKT-067-06	JA2828
Exh 5; "Standards for Educational and Psychological Testing" Sales Report, 1999 Edition AERA-DKT-067-07	JA2832
Exh 6; AERA Standards for Educational and Psychological Testing Statement of Revenue and Expenses; (FY2000 - 12/31/14) AERA-DKT-067-08 (Material Under Seal).....	JA2834
Exh 7; Standards for Educational and Psychological Testing (Fund Balance Report) AERA-DKT-067-09 (Material Under Seal).....	JA2836
Exh 8; American Psychological Association - APA Membership Statistics AERA-DKT-067-10	JA2838
Exh 9; Standards for Educational & Psychological Testing (2014 Edition) AERA-DKT-067-11	JA2848
"Exh 10; Briefing Room - Remarks by the President in State of the Union Address" AERA-DKT-067-12	JA2852
Exh 11; Everything You Need to Know: Waivers, Flexibility, and Reforming No Child Left Behind AERA-DKT-067-13	JA2870
Exh 12; National resolution against high-stakes tests released AERA-DKT-067-14	JA2876

TABLE OF CONTENTS
(Continued)

Page

VOLUME VI:

Exh 13; FairTest - Resistance to High Stakes Testing Spreads AERA-DKT-067-15	JA2882
Exh 14; California Intellectual Property Laws, 2015 Edition; Publisher: Matthew Bender AERA-DKT-067-16	JA2886
Exh 15; Amazon - Code of Federal Regulations, Title 38, Pensions, Bonuses, and Veterans' Relief, Pt. 0-17, Revised as of July 1, 2015 AERA-DKT-067-17	JA2889
Exh 16; Barnes & Noble Classics - web search results AERA-DKT-067-18	JA2893
Exh 2; Rule 30(b)(6) Deposition of Dianne L. Schneider, dated 04/23/15 (excerpts) AERA-DKT-068-06 (Material Under Seal)	JA2897
Exh 3; Rule 30(b)(6) Deposition of AERA, APA, NCME, representative Marianne Ernesto, dated 04/29/15 (excerpts) AERA-DKT-068-07 (Material Under Seal)	JA2903
Exh 5; Deposition of Felice J. Levine, dated 05/04/15 (excerpts) AERA-DKT-068-09 (Material Under Seal)	JA2927
Exh 6; Deposition of Laress L. Wise; dated 05/11/15 (excerpts) AERA-DKT-068-10 (Material Under Seal)	JA2956
Exh 8; Deposition of Kurt F. Geisinger, dated 09/10/15 (excerpts) AERA-DKT-068-11 (Material Under Seal)	JA2962
Exh 11; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 04/24/14 AERA-DKT-068-12 (Material Under Seal)	JA2990

TABLE OF CONTENTS
(Continued)

	Page
Exh 13; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 09/10/14 AERA-DKT-068-14 (Material Under Seal).....	JA2993
Exh 14; Copyright Assignment, from Leonard S. Feldt, to AERA, APA, NCME, dated 12/12/14 AERA-DKT-068-15 (Material Under Seal).....	JA2996
Exh 15; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 09/08/14 AERA-DKT-068-16 (Material Under Seal).....	JA3001
Exh 17; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 04/21/14 AERA-DKT-068-17 (Material Under Seal).....	JA3004
Exh 18; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 04/22/14 AERA-DKT-068-18 (Material Under Seal).....	JA3007
Exh 19; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 04/21/14 AERA-DKT-068-19 (Material Under Seal).....	JA3012
Exh 20; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 10/16/14 AERA-DKT-068-20 (Material Under Seal).....	JA3015
Exh 21; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 04/28/14 AERA-DKT-068-21 (Material Under Seal).....	JA3019
Exh 22; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 09/08/14 AERA-DKT-068-22 (Material Under Seal).....	JA3022

TABLE OF CONTENTS
(Continued)

	Page
Exh 23; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 10/21/14 AERA-DKT-068-23 (Material Under Seal).....	JA3025
Exh 24; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 04/23/14 AERA-DKT-068-24 (Material Under Seal).....	JA3028
Exh 25; Correspondence from AERA, Am. Psychological Assoc., NCME dated 04/21/14; executed 04/24/14 AERA-DKT-068-25 (Material Under Seal).....	JA3031
Exh 26; Copyright Assignment, from Charlie Spielberger to AERA, APA, NCME dated 12/30/14 AERA-DKT-068-26 (Material Under Seal).....	JA3034
Exh 28; Copyright Certificate of Registration - Standards for Educational and Psychological Testing, dated 02/25/14 AERA-DKT-068-28 (Material Under Seal).....	JA3038
Exh 29; Standards for Educational and Psychological Testing AERA-DKT-068-29 (Material Under Seal).....	JA3041
Exh 30; APA - Professional Standards to Ensure the Fair and Appropriate Use of Testing in High-Stakes Educational Decisions AERA-DKT-068-30 (Material Under Seal).....	JA3050
Exh 32; Highlights of APA's Involvement in Educational Testing Provisions of the "No Child Left Behind Act" AERA-DKT-068-31 (Material Under Seal).....	JA3054
Exh 33; Correspondence to Elliott Eisner and Ronald A. Berk from Frank Farley, APA, dated 01/05/93 AERA-DKT-068-32 (Material Under Seal).....	JA3058
"Exh 38; American Educational Research Association Standards of Educational and Psychological Testing Statement of Revenue and Expenses (FY2000 -	

TABLE OF CONTENTS
(Continued)

	Page
12/31/2013)" AERA-DKT-068-34 (Material Under Seal).....	JA3063
"Exh 41; American Educational Research Association Standards of Educational and Psychological Testing Statement of Revenue and Expenses (FY2000 - 12/31/2014)" AERA-DKT-068-35 (Material Under Seal).....	JA3065
Declaration of Carl Malamud in Support of Public.Resource.Org's Motion for Summary Judgment, dated 01/21/16 AERA-DKT-069-05	JA3067
Exh 10; Copyright Registration - Standards for Educational and Psychological Testing, dated 12/08/99 AERA-DKT-070-10	JA3078
Exh 39; Standards for Educational and Psychological Testing Sales Report, 1999 Edition AERA-DKT-070-38	JA3081
Exh 40; Standards for Educational and Psychological Testing Sales Report AERA-DKT-070-39	JA3083
Exh 44; Monitor on Psychology - advertisement for Standards for Educational and Psychological Testing AERA-DKT-070-43	JA3085
Exh 45; Table of Contents - Standards for Educational and Psychological Testing AERA-DKT-070-44	JA3087
Exh 46; New! Revised! Expanded! Standards for Educational and Psychological AERA-DKT-070-45	JA3098
Exh 47; AERA web listing - Standards for Educational & Psychologic (2014 Edition) AERA-DKT-070-46	JA3102

TABLE OF CONTENTS
(Continued)

	Page
Exh 48; AERA web listing - Standards for Educational & Psychologic (2014 Edition) AERA-DKT-070-47	JA3110
Exh 51; Expert Report of James R. Fruchterman, dated 06/13/15 AERA-DKT-070-50	JA3114
Exh 52; SIOP article - OCR Issues Draft Guide on Disparate Impact in Educational Testing, Wayne Camara, The College Board AERA-DKT-070-51	JA3241
Exh 65; National Archives - Incorporation by Reference AERA-DKT-070-64	JA3246
Plaintiff's Reply in Further Support of its Motion for Summary Judgement and in Opposition to Defendant's Motion for Summary Judgment, dated 02/18/16 AERA-DKT-089.....	JA3251
Exh 83; Report of the Advisory Commission on Accessible Instructional Materials in Postsecondary Education for Students with Disabilities, 12/06/11 AERA-DKT-099-13	JA3253
Order, re: Defendant's Motion to Strike the Geisinger Declaration AERA-DKT-115.....	JA3256
Transcript of Motions Hearing before the Honorable Tanya S. Chutkan, dated 09/12/16 AERA-DKT-116.....	JA3259
Memorandum Opinion, Dated 02/02/17 AERA-DKT-117.....	JA3401
Order, Dated 02/02/17 AERA-DKT-118.....	JA3456
Plaintiff's Motion for Clarification of the Court's Order dated February 2, 2017, dated 02/10/17 AERA-DKT-119.....	JA3457

**TABLE OF CONTENTS
(Continued)**

	Page
Notice of Appeal by Defendant-Counterclaimant Public.Resource.Org, Inc., dated 02/17/17	
AERA-DKT-120.....	JA3461

CERTIFICATE OF SERVICE

I, hereby certify that on January 31, 2018, I electronically filed the foregoing **Appendix** with the Clerk of the United States Court of Appeals for the District of Columbia Circuit by using the appellate CM/ECF system. I certify that all participants in the case are registered CM/ECF users and that service will be accomplished by the appellate CM/ECF system.

By: /s/ Andrew P. Bridges

Andrew P. Bridges (admitted)

abridges@fenwick.com

Matthew B. Becker (admitted)

mbecker@fenwick.com

FENWICK & WEST LLP

555 California Street, 12th Floor

San Francisco, CA 94104

Telephone: (415) 875-2300

Facsimile: (415) 281-1350

Attorneys for Appellant

Public.Resource.Org, Inc.

EXHIBIT V-2

Case No. 1:14-cv-00857-TSC-DAR

9. TESTING INDIVIDUALS OF DIVERSE LINGUISTIC BACKGROUNDS

Background

For all test takers, any test that employs language is, in part, a measure of their language skills. This is of particular concern for test takers whose first language is not the language of the test. Test use with individuals who have not sufficiently acquired the language of the test may introduce construct-irrelevant components to the testing process. In such instances, test results may not reflect accurately the qualities and competencies intended to be measured. In addition, language differences are almost always associated with concomitant cultural differences that need to be taken into account when tests are used with individuals whose dominant language is different from that of the test. Whether a certain dialect of a language should be considered a different language cannot be resolved here, although some aspects of the present discussion are relevant to the debate. In either case, special attention to issues related to language and culture may be needed when developing, administering, scoring, and interpreting test scores and making decisions based on test scores. Language proficiency tests, if appropriately designed and used, are an obvious exception to this concern because they are intended to measure familiarity with the language of the test as required in educational and other settings.

Individuals who are bilingual can vary considerably in their ability to speak, write, comprehend aurally, and read in each language. These abilities are affected by the social or functional situations of communication. Some people develop socially and culturally acceptable ways of speaking that combine two or more languages simultaneously. Other individuals familiar with two languages may perform more slowly, less efficiently, and at times less accurately on prob-

lem-solving tasks that are administered in the less familiar language. Language dominance is not necessarily an indicator of language competence in taking a test, and some accommodation may be necessary even when administering the test in the more familiar language. Therefore it is important to consider language background in developing, selecting, and administering tests and in interpreting test performance. Consequently, for example, test norms based on native speakers of English either should not be used with individuals whose first language is not English or such individuals' test results should be interpreted as reflecting in part current level of English proficiency rather than ability, potential, aptitude or personality characteristics or symptomatology. In cases where a language-oriented test is inappropriate due to the test takers' limited proficiency in that language, a non-verbal test may be a suitable alternative.

Where effective job performance requires the ability to communicate in the language of the test, persons who do not have adequate proficiency in that language may perform poorly on the test, on the job, or both. In that case, the tests used for prediction of future job performance appropriately would be administered in the language of the job, as long as the language level needed for the test did not exceed the level needed to meet work requirements. Test users should understand that poor test performance, as well as poor job performance, may result from poor language proficiency rather than other deficiencies.

Many issues addressed in this chapter are also relevant to testing individuals who have unique linguistic characteristics due to disabilities such as deafness and/or blindness. For example, issues regarding test translation and adaptation are applicable to American Sign Language (ASL) versions of traditional tests. It should be noted, however, that ASL is

TESTING INDIVIDUALS OF DIVERSE LINGUISTIC BACKGROUNDS / PART II

not only a different language but is also a different mode of communication. Also, individuals with disabilities may require modifications in test administration procedures similar to those required by non-native speakers. A more specific discussion of testing individuals with disabilities is provided in chapter 10.

Issues discussed in earlier chapters, in particular chapters 1-5, including validity of test score inferences, test reliability, and test development and administration are germane to this chapter. The present chapter extends these discussions, emphasizing the importance of recognizing the possible impact of language abilities and skills on test performance. There may be legal requirements relevant to the testing of individuals with different language backgrounds. The standards in this chapter are intended to be applied in a manner consistent with those requirements.

Test Translation, Adaptation, and Modification

Testing test takers in their primary language may be necessary in order to draw valid inferences based on their test scores. Thus, language modifications are often needed. Translating a test to the primary language represents one such modification. However, a number of hazards need to be avoided when doing this sort of translation. One cannot simply assume that such a translation produces a version of the test that is equivalent in content, difficulty level, reliability, and validity to the original untranslated version. Further, one cannot assume that test takers' relevant acculturation experiences are comparable across the two versions. Also, many words have different frequency rates or difficulty levels in various languages. Therefore, words in two languages that appear to be close in meaning may differ significantly in ways that seriously impact the translated test for the intended test use. Additionally, the test content of the translated version may not be equivalent to

that of the original version. For example, a test of reading skills in language A that is translated to serve as a test of reading skills in language B may include content not equally meaningful or appropriate for people who read only language B.

For the purposes of test translation and adaptation for use with test takers whose first language is not the language of the test, back translation is not recommended as a stand-alone procedure. It may provide an artificial similarity of meaning across languages but not the best version in the new language. In most situations, an iterative process more akin to test development and validation is suggested to ensure that similar constructs are measured across versions. When test forms in two or more languages are developed concurrently, it is generally desirable that some items originate in each of the languages involved. The decision as to whether to use the standard original language test or an adapted version is a complex matter. Issues that may have an impact on this decision are discussed in the next section.

Other strategies of test modification may be appropriate when the test taker's primary language is not the language of the test. These include modifying aspects of the test or the test administration procedure such as the presentation format, the response format, the time allowed to complete the test, the test setting (individual administration instead of group testing), and the use of only those portions of the test that are appropriate for the level of language proficiency of the test taker. If modifications are made to the presentation or response format of the test, it may sometimes be appropriate for the modified test to be field tested with an adequate population sample prior to use with its intended population.

Issues of Equivalence

The term *equivalence*, as used here, refers to the degree to which test scores can be used to make comparable inferences for different

PART II / TESTING INDIVIDUALS OF DIVERSE LINGUISTIC BACKGROUNDS

examinees. When tests are designed for and used with linguistically homogeneous populations, issues of equivalence are relatively straightforward (for example, see chapters 1 and 4). If an individual examinee can be demonstrated to belong to the population for which the test was designed, then adhering to standard procedures of test administration and interpretation is expected to lead to reliable and valid inferences based on the examinee's test score. When a test is intended for use with test takers who differ linguistically from those for whom the test was designed, establishing equivalence poses a greater challenge. In general, the linguistic and cultural characteristics of the intended examinee population should be reflected in examinee samples used throughout the processes of test design, validation, and norming. At each of these stages of test development and standardization, distinct linguistic groups should receive the same level of specific attention. The inclusion of proportional representation of linguistic subgroups in aggregate standardization and validation samples may be insufficient to assure equivalence across linguistic groups.

Issues associated with construct equivalence are perhaps most fundamental. One may question whether the test score for a particular individual represents that individual's standing with respect to the same construct as is measured in the target population. For example, among non-native speakers of the language of the test, one may not know whether a test designed to measure primarily academic achievement becomes in whole or in part a measure of proficiency in the language of the test. There are several psychometric techniques that can be used to determine the equivalence of constructs across groups, including confirmatory factor analysis, analysis of data contained in multi-method-multitrait matrices and the equivalence of responsiveness of the groups to experimental manipulations. These tech-

niques may be supplemented with logical analyses of the results based on knowledge of the linguistic characteristics of the test taker's population of origin.

Other types of equivalence also need to be considered when testing individuals from different linguistic backgrounds. Functional equivalence addresses the question of whether similar activities or behaviors measured by a test have the same meaning in different cultural or linguistic groups. Translation equivalence requires that the translated or adapted test be comparable in content to the original test; it was addressed above in the discussion of test translation and adaptation. Metric equivalence concerns the issue of whether scores from the same test administered in different languages have comparable psychometric properties. For example, with metric equivalence, a score of 50 on test X in language A is interpretable in the same way as a score of 50 on test X in language B. In general, metric equivalence will be limited to particular contexts, examinee groups, and types of interpretations.

Language Proficiency Testing

Consideration of relevant within-linguistic group differences is crucial in determining appropriate test interpretation and decision making in educational programs and in some professional applications of individualized tests. For example, individuals whose first language is not the language of the test may vary considerably in their proficiency along a continuum from those who have no knowledge of the language of the test to those who are fluent in it and knowledgeable of the corresponding culture. Further, a demographic proxy such as Mexican or German is likely to prove insufficient in determining the language of test administration because members of the same cultural group may vary widely in their degree of acculturation, proficiency in the language of the test, familiarity with words and syntax in their native languages,

TESTING INDIVIDUALS OF DIVERSE LINGUISTIC BACKGROUNDS / PART II

educational background, familiarity with tests and test-taking skills, and other factors that may significantly affect the reliability and validity of inferences drawn from test scores. Thus, it is essential that individual differences that may affect test performance be taken into account when testing individuals of differing linguistic backgrounds.

The need exists to consider both language dominance and language proficiency. Standardized tests that assess multiple domains in a given language can be helpful in determining language dominance and proficiency. The person conducting the testing first should obtain information about the language in which the examinee is dominant (i.e., the preferred or salient language). Following this determination of dominance, the examinee's level of proficiency in the dominant language should be established. If the languages are similarly dominant, then proficiency should be established for both (or all) languages. Then the test should be administered in the most proficient language if available (unless the purpose of the testing is to determine proficiency in the language of the test). However, testing individuals in their dominant language alone is no panacea because, as suggested above, a bilingual individual's two languages are likely to be specialized by domain (e.g., the first language is used in the context of home, religious practices, and native culture, whereas the second language is used in the context of school, work, television, and mainstream culture). Thus, a test in either language by itself will likely measure some domains and miss out on others. In such situations, testing in both languages (i.e., the dominant language and the language in which the test taker is most proficient) may be necessary, provided appropriate tests are available. If assessment in both languages is carried out, careful consideration should be given to the possibility of order effects.

Because students are expected to acquire proficiency in the language used in schools that is appropriate to their ages and educational levels, tests suitable for assessing their progress in that language are needed. For example, some tests, especially paper-and-pencil measures, that are prepared for students of English as a foreign language may not be particularly useful if they place insufficient emphasis on the assessment of important listening and speaking skills. Measures of competency in all relevant English language skills (e.g., communicative competence, literacy, grammar, pronunciation, and comprehension) are likely to be most valuable in the school context.

Observing students' speech in naturalistic situations can provide additional information about their proficiency in a language. However, findings from naturalistic observations may not be sufficient to judge students' ability to function in that language in formal, academically oriented situations (e.g., classrooms). For example, it is not appropriate to base judgments of a child's ability to benefit from instruction in one language solely on language fluency observed in speech use on the playground. Nor is it appropriate to base judgments of a person's ability to perform a job on assessments of formal language usage, if formal language usage is not linked to job performance.

In general, there are special difficulties attendant upon the use of a test with individuals who have not had an adequate opportunity to learn the language used by the test. When a test is used to inform a decision process that has a broad impact, it may be important for the test user to review the test itself and to consider the possible use of alternative information-gathering tools (e.g., additional tests, sources of observational information, modified forms of the chosen test) to ensure that the information obtained is adequate to the intended purpose. Reviews of this kind may sometimes reveal the need

PART II / TESTING INDIVIDUALS OF DIVERSE LINGUISTIC BACKGROUNDS

to create a formal adaptation of a test or to develop a new test that is suitable for the specific linguistic characteristics of the individuals being tested.

Testing Bilingual Individuals

Test use with examinees who are bilingual also poses special challenges. An individual who knows two languages may not test well in either language. As an example, children from homes where parents speak Spanish may be able to understand Spanish but express themselves best in English. In addition, some persons who are bilingual use their native language in most social situations and use English primarily for academic and work-related activities; the use of one or both languages depends on the nature of the situation. As another example, proficiencies in conversational English and written English can often differ. Non-native English speakers who may give the impression of being fluent in conversational English may not be competent in taking tests that require English literacy skills. Thus, an understanding of an individual's type and degree of bilingualism is important to proper test use.

Administration and Examiner Variables

When an examinee cannot be assumed to belong to the cultural or linguistic population upon which the test was standardized, then use of standardized administration procedures may not provide a comparable administration of the test for that examinee. In this situation, the fundamental principle of sound practice is that examinees, regardless of background, should be provided with an adequate opportunity to complete the test and demonstrate their level of competence on the attributes the test is intended to measure. There may be, however, complex interactions among examiner, examinee, and situational variables

that require careful attention on the part of the practitioner administering the test. Factors that may affect the performance of the examinee include the cultural and linguistic background of both the examiner and examinee; the gender and testing style of the examiner; the level of acculturation of the examinee and examiner; whether the test is administered in the original language of the test, the examinee's primary language, or whether both languages are used (and if so in what order); the time limits of the testing; and whether a bilingual interpreter is used.

Use of Interpreters in Testing

Ideally, when an adequately translated version of the test or a suitable nonverbal test is unavailable, assessment of individuals with limited proficiency in the language of the test should be conducted by a professionally trained bilingual examiner. The bilingual examiner should be proficient in the language of the examinee at the level of a professional trained in that language. When a bilingual examiner is not available, an alternative is to use an interpreter in the testing process and administer the test in the examinee's native language. Although a commonly used procedure, this practice has some inherent difficulties. For example, there may be a lack of linguistic and cultural equivalence between the translation and the original test, the translator or the interpreter may not be adequately trained to work in the testing situation, and representative norms may not be available to score and interpret the test results appropriately. These difficulties may pose significant threats to the validity of inferences based on test results.

When the need for an interpreter arises for a particular testing situation, it is important to obtain a fully qualified interpreter to assist the examiner in administering the test. The most important consideration in testing with the services of an interpreter is the inter-

TESTING INDIVIDUALS OF DIVERSE LINGUISTIC BACKGROUNDS / PART II

preter's ability and preparedness in carrying out the required duties during testing. The interpreter obviously needs to be fluent in both the language of the test and the examinee's native language and have general familiarity with the process of translating. To be effective, the interpreter also needs to have a basic understanding of the process of psychological and educational assessment, including the importance of following standardized procedures, the importance of accurately conveying to the examiner an examinee's actual responses, and the role and responsibilities of the interpreter in testing. Additionally, it is inappropriate for the interpreter to have any prior personal relationship with the test taker that is likely to jeopardize the objectivity of the test administration. However, in small linguistic or cultural communities, speakers of the alternate languages are often known to each other. Therefore, in such cases, it is the responsibility of the test user or examiner to ensure that the interpreter has received adequate instruction in the principles of objective test administration and to assess preexisting biases so that test interpretations can take such factors into account. If clear biases are evident and cannot be ameliorated, then the examiner should make arrangements to obtain another interpreter.

Whenever proficiency in the language of the test is essential to job performance, use of a translator to assist a candidate with licensure, certification, or civil service examinations should be permitted only when it will not compromise standards designed to protect public health, safety, and welfare. When a translator is permitted, it also is essential that the candidate not receive help interpreting the content of the test or any other assistance that would compromise the integrity of the licensure or certification decision. Creation of audio tapes that enable a candidate to listen to each question being read in the language of the test may be more appropriate when such an accommodation is justified.

In educational and psychological testing, it may be appropriate for an interpreter to become familiar with all details of test content and administration prior to the testing. Also, time needs to be provided for the interpreter to translate test instructions and items, if necessary. In psychological testing, it is often desirable for the examiner to demonstrate for the interpreter how certain test items are administered and explain what to expect during testing. In addition, it is important that, prior to the testing, the examiner and the interpreter become familiar with each other's style of speaking and the speed at which they work. Immediately prior to the assessment, the role of the interpreter needs to be explained clearly to the examinee. It is essential that the interpreter make all efforts to provide accurate information in translation. The interpreter must reflect a professional attitude and maintain objectivity throughout the testing process (e.g., not interject subjective opinions, not give cues to the examinee). Once the testing is completed, the examiner is responsible for reviewing the test responses with the assistance of the interpreter. Responses that are difficult to interpret (e.g., vocabulary words), nontest behaviors that might have special meanings (e.g., body language), as well as language factors (e.g., mixed use of two languages) and cultural factors that might have an effect on testing results need to be discussed fully. This information is to be used then by the examiner in carefully evaluating the test results and drawing inferences from the results.

Cultural Differences and Individual Testing

Linguistic behavior that may appear eccentric or be judged to be less appropriate in one culture may be seen as more appropriate in another culture and may need to be taken into account during the testing process. For example, children or adults from some cul-

tures may be reluctant to speak in elaborate language to adults or people in higher status roles and instead may be encouraged to speak to such persons only in response to specific questions or with formulaic utterances. Thus, when tested, such test takers may respond to an examiner probing for elaborate speech with only short phrases or by shrugging their shoulders. Interpretations of scores resulting from such testing may prove to be inaccurate if this tendency is not properly taken into consideration. At the same time, the examiner should not presume that their reticence is necessarily a cultural characteristic. Additional information (e.g., prior observations or a family member's consultation) may be needed to discuss the extent of culture's possible influence on linguistic performance.

The values associated with the nature and degree of verbal output also may differ across cultures. One cultural group may judge verbosity or rapid speech as rude, whereas another may regard those speech patterns as indications of high mental ability or friendliness. An individual from one culture who is evaluated with values appropriate to another culture may be considered taciturn, withdrawn, or of low mental ability. Resulting interpretations and prescriptions of treatment may be invalid and potentially harmful to the individual being tested.

Standard 9.1

Testing practice should be designed to reduce threats to the reliability and validity of test score inferences that may arise from language differences.

Comment: Some tests are inappropriate for use with individuals whose knowledge of the language of the test is questionable. Assessment methods together with careful professional judgment are required to determine when language differences are relevant. Test users can judge how best to address this standard in a particular testing situation.

Standard 9.2

When credible research evidence reports that test scores differ in meaning across subgroups of linguistically diverse test takers, then to the extent feasible, test developers should collect for each linguistic subgroup studied the same form of validity evidence collected for the examinee population as a whole.

Comment: Linguistic subgroups may be found to differ with respect to appropriateness of test content, the internal structure of their test responses, the relation of their test scores to other variables, or the response processes employed by individual examinees. Any such findings need to receive due consideration in the interpretation and use of scores as well as in test revisions. There may also be legal or regulatory requirements to collect subgroup validity evidence. Not all forms of evidence can be examined separately for members of all linguistic groups. The validity argument may rely on existing research literature, for example, and such literature may not be available for some populations. For some kinds of evidence, separate linguistic subgroup analyses may not be feasible due to the limited number of cases available. Data may sometimes be accumulated so that these

STANDARDS

TESTING INDIVIDUALS OF DIVERSE LINGUISTIC BACKGROUNDS / PART II

analyses can be performed after the test has been in use for a period of time. It is important to note that this standard calls for more than representativeness in the selection of samples used for validation or norming studies. Rather, it calls for separate, parallel analyses of data for members of different linguistic groups, sample sizes permitting. If a test is being used while such data are being collected, then cautionary statements are in order regarding the limitations of interpretations based on test scores.

Standard 9.3

When testing an examinee proficient in two or more languages for which the test is available, the examinee's relative language proficiencies should be determined. The test generally should be administered in the test taker's most proficient language, unless proficiency in the less proficient language is part of the assessment.

Comment: Unless the purpose of the testing is to determine proficiency in a particular language or the level of language proficiency required for the test is a work requirement, test users need to take into account the linguistic characteristics of examinees who are bilingual or use multiple languages. This may require the sole use of one language or use of multiple languages in order to minimize the introduction of construct-irrelevant components to the measurement process. For example, in educational settings, testing in both the language used in school and the native language of the examinee may be necessary in order to determine the optimal kind of instruction required by the examinee. Professional judgement needs to be used to determine the most appropriate procedures for establishing relative language proficiencies. Such procedures may range from self-identification by examinees through formal proficiency testing.

Standard 9.4

Linguistic modifications recommended by test publishers, as well as the rationale for the modifications, should be described in detail in the test manual.

Comment: Linguistic modifications may be recommended for the original test in the primary language or for an adapted version in a secondary language, or both. In any case, the test manual should provide appropriate information regarding the recommended modifications, their rationales, and the appropriate use of scores obtained using these linguistic modifications.

Standard 9.5

When there is credible evidence of score comparability across regular and modified tests or administrations, no flag should be attached to a score. When such evidence is lacking, specific information about the nature of the modification should be provided, if permitted by law, to assist test users properly to interpret and act on test scores.

Comment: The inclusion of a flag on a test score where a linguistic modification was provided may conflict with legal and social policy goals promoting fairness in the treatment of individuals of diverse linguistic backgrounds. If a score from a modified administration is comparable to a score from a nonmodified administration, there is no need for a flag. Similarly, if a modification is provided for which there is no reasonable basis for believing that the modification would affect score comparability, there is no need for a flag. Further, reporting practices that use asterisks or other non-specific symbols to indicate that a test's administration has been modified provide little useful information to test users.

PART II / TESTING INDIVIDUALS OF DIVERSE LINGUISTIC BACKGROUNDS

STANDARDS

Standard 9.6

When a test is recommended for use with linguistically diverse test takers, test developers and publishers should provide the information necessary for appropriate test use and interpretation.

Comment: Test developers should include in test manuals and in instructions for score interpretation explicit statements about the applicability of the test with individuals who are not native speakers of the original language of the test. However, it should be recognized that test developers and publishers seldom will find it feasible to conduct studies specific to the large number of linguistic groups found in certain countries.

Standard 9.7

When a test is translated from one language to another, the methods used in establishing the adequacy of the translation should be described, and empirical and logical evidence should be provided for score reliability and the validity of the translated test's score inferences for the uses intended in the linguistic groups to be tested.

Comment: For example, if a test is translated into Spanish for use with Mexican, Puerto Rican, Cuban, Central American, and Spanish populations, score reliability and the validity of test score inferences should be established with members of each of these groups separately where feasible. In addition, the test translation methods used need to be described in detail.

Standard 9.8

In employment and credentialing testing, the proficiency level required in the language of the test should not exceed that appropriate to the relevant occupation or profession.

Comment: Many occupations and professions require a suitable facility in the language of the test. In such cases, a test that is used as a part of selection, advancement, or credentialing may appropriately reflect that aspect of performance. However, the level of language proficiency required on the test should be no greater than the level needed to meet work requirements. Similarly, the modality in which language proficiency is assessed should be comparable to that on the job. For example, if the job requires only that employees understand verbal instructions in the language used on the job, it would be inappropriate for a selection test to require proficiency in reading and writing that particular language.

Standard 9.9

When multiple language versions of a test are intended to be comparable, test developers should report evidence of test comparability.

Comment: Evidence of test comparability may include but is not limited to evidence that the different language versions measure equivalent or similar constructs, and that score reliability and the validity of inferences from scores from the two versions are comparable.

Standard 9.10

Inferences about test takers' general language proficiency should be based on tests that measure a range of language features, and not on a single linguistic skill.

Comment: For example, a multiple-choice, pencil-and-paper test of vocabulary does not indicate how well a person understands the language when spoken nor how well the person speaks the language. However, the test score might be helpful in determining how well a person understands some aspects of the written language. In making educational

STANDARDS

TESTING INDIVIDUALS OF DIVERSE LINGUISTIC BACKGROUNDS / PART II

placement decisions, a more complete range of communicative abilities (e.g., word knowledge, syntax) will typically need to be assessed.

Standard 9.11

When an interpreter is used in testing, the interpreter should be fluent in both the language of the test and the examinee's native language, should have expertise in translating, and should have a basic understanding of the assessment process.

Comment: Although individuals with limited proficiency in the language of the test should ideally be tested by professionally trained bilingual examiners, the use of an interpreter may be necessary in some situations. If an interpreter is required, the professional examiner is responsible for ensuring that the interpreter has the appropriate qualifications, experience, and preparation to assist appropriately in the administration of the test. It is necessary for the interpreter to understand the importance of following standardized procedures, how testing is conducted typically, the importance of accurately conveying to the examiner an examinee's actual responses, and the role and responsibilities of the interpreter in testing.

10. TESTING INDIVIDUALS WITH DISABILITIES

Background

With the advancement of scientific knowledge, medical practices, and social policies, increasing numbers of individuals with disabilities are participating more fully in educational, employment, and social activities. This increased participation has resulted in a greater need for the testing and assessment of individuals with disabilities for a variety of purposes. Individuals with disabilities are defined as persons possessing a physical, mental, or developmental impairment that substantially limits one or more of their major life activities. Although the *Standards* focus on technical and professional issues regarding the testing of individuals with disabilities, test developers and users are encouraged to become familiar with federal, state, and local laws, and court and administrative rulings that regulate the testing and assessment of individuals with disabilities.

Tests are administered to individuals with disabilities in various settings and for diverse purposes. For example, tests are used for diagnostic purposes to determine the existence and nature of a test taker's disabilities. Testing is also conducted for prescriptive purposes to determine intervention plans. In addition, tests are administered to persons who have been diagnosed with identified disabilities for educational and employment purposes to make placement, selection, or other similar decisions, or for monitoring performance as a tool for educational accountability. These uses of tests for persons with disabilities occur in a variety of contexts including school, clinical, counseling, forensic, employment, and credentialing.

Issues Regarding Accommodation When Testing Individuals With Disabilities

A major issue when testing individuals with disabilities concerns the use of accommoda-

tions, modifications, or adaptations. The purpose of these accommodations or modifications is to minimize the impact of test-taker attributes that are not relevant to the construct that is the primary focus of the assessment. The terms *accommodation* and *modification* have varying connotations in different subfields. Here *accommodation* is used as the general term for any action taken in response to a determination that an individual's disability requires a departure from established testing protocol. Depending on circumstances, such accommodation may include modification of test administration processes or modification of test content. No connotation that modification implies a change in the construct(s) being measured is intended.

A standardized test that has been designed for use with the general population may be inappropriate for use for individuals with specific disabilities if the test requires the use of sensory, motor, language, or psychological skills that are affected by the disability and that are not relevant to the focal construct. For example, a person who is blind may read only in Braille format, and an individual with hemiplegia may be unable to hold a pencil and thus would have difficulty completing a standard written exam. In addition, some individuals with disabilities may possess other attendant characteristics (e.g., a person with a physical disability may fatigue easily), causing them to be further challenged by some standardized testing situations. In these examples, if reading, use of a pencil, and fatigue are incidental to the construct intended to be measured by the test, modifications of tests and test administration procedures may be necessary for an accurate assessment.

Note also that accommodations are not needed or appropriate under a variety of circumstances. First, the disability may, in fact, be directly relevant to the focal construct. For example, no accommodation is appropriate for a person who is completely blind if the

TESTING INDIVIDUALS WITH DISABILITIES / PART II

test is designed to measure visual spatial ability. Similarly, in employment testing it would be inappropriate to make test modifications if the test is designed to assess essential skills required for the job and the modifications would fundamentally alter the constructs being measured. Second, an accommodation for a particular disability is inappropriate when the purpose of a test is to diagnose the presence and degree of that disability. For example, allowing extra time on a timed test to assess the existence of a specific learning disability would make it very difficult to determine if a processing difficulty actually exists. Third, it is important to note that not all individuals with disabilities require special provisions when taking all tests. Many individuals have disabilities that would not influence their performance on a particular test, and hence no modification is needed.

Professional judgment necessarily plays a substantial role in decisions about test accommodations. Judgment comes into play in determining whether a particular individual needs accommodation and the nature and extent of such accommodation. In some circumstances, individuals with disabilities request testing accommodations and provide appropriate documentation in support of the request. Generally the request is reviewed by the agency sponsoring the assessment or an outside source knowledgeable about the assessment process and the type of disability. In either case, a conclusion is drawn as to what constitutes reasonable accommodation. Disagreement may arise between the accommodation requested by an individual with a disability and the granted accommodation. In these situations, and to the extent permitted by law, the overarching concern is the validity of the inference made from the score on the modified test: fairness to all parties is best served by a decision about test modification that results in the most accurate measure possible of the construct of interest. The role of professional judgment is further complicated by the fact that empirical research on test accommodations is often lacking.

When modifying tests it is also important to recognize that individuals with the same type of disability may differ considerably in their need for accommodation. A central consideration in determining a test modification for a disability is to recognize that the modifications should be tailored directly to the specific needs of individual test takers. As an example, it would be incorrect to make the assumption that all individuals with visual impairments would be successfully accommodated by providing testing materials in Braille format. Depending on the extent of the disability, it may be more appropriate for some individuals to receive testing materials written in large print, while others might need a tape cassette or reader.

As test modifications involve altering some aspect of a test originally developed for use with a target population, it is important to recognize that making these alterations has the potential to affect the psychometric qualities of the test. There have been few empirical investigations into the effects of various accommodations on the reliability of test scores or the validity of inferences drawn from modified tests. Due to a number of practical limitations (e.g., small sample size, nonrandom selection of test takers with disabilities), there is no precise, technical solution available for equating modified tests to the original form of these tests. Thus it is difficult to compare scores from a test modified for persons with disabilities with scores from the original test.

Modifications designed to accommodate persons with disabilities also may change the construct measured by the test, or the extent to which it is fully measured. For example, a test of oral comprehension may become a test of reading comprehension when administered in written format to a person who is deaf or hard of hearing. Such a change in test administration may alter the construct being measured by the original test. When this occurs, the scores on the standard and modified versions of the test will not have the same meaning. Similarly, modification of test administration may also

PART II / TESTING INDIVIDUALS WITH DISABILITIES

alter the predictive value of test scores. For example, when a speed test is administered with relaxed time requirements to a person with a disability, the relationship of test scores to criteria such as job performance may be affected. Appropriate professional judgment should be exercised in interpreting and using scores on modified tests.

Some modified tests, with accompanying research to support the appropriate modifications, have been available for a number of years. Although the development of tests and testing procedures for individuals with disabilities is encouraged by the *Standards*, it should be noted that all relevant individual standards given elsewhere in this document are fully applicable to the testing applications and modifications or accommodations considered in this chapter. Issues of validity and reliability are critical whenever modifications or accommodations occur.

Strategies of Test Modification

A variety of test modification strategies have been implemented in various settings to accommodate the needs of test takers with disabilities. Some require modifying test administration procedures (e.g., instructions, response format) while others alter test medium, timing, settings, or content. Depending on the nature and extent of the disability, one or more test modification procedures may be appropriate for a particular individual. The listing here of a variety of modification strategies should not suggest that the full array of strategies is routinely available or appropriate; the decision to modify rests on a determination that modification is needed to make valid inferences about the individual's standing on the construct in question.

MODIFYING PRESENTATION FORMAT

One modification option is to alter the medium used to present the test instructions and items to the test takers. For example, a test booklet may be produced in Braille or large print for individuals with visual impairments. When tests are computer-administered,

larger fonts or oversized computer screens may be used. Individuals with a hearing disability may receive test instructions through the use of sign communication or writing.

MODIFYING RESPONSE FORMAT

Modifications also can be made to allow individuals with disabilities to respond to test items using their preferred communication modality. For example, an individual with severe language deficits might be allowed to point to the preferred response. A test taker who cannot manually record answers to test items or questions may be assisted by an aide who would mark the answer. Other ways of obtaining a response include having the respondent use a tape recorder, a computer keyboard, or a Braillewriter.

MODIFYING TIMING

Another modification available is to alter the timing of tests. This may include extended time to complete the test, more breaks during testing, or extended testing sessions over several days. Many national testing programs (e.g., achievement, certification) allow persons with disabilities additional time to take the test. Reading Braille, using a cassette recorder, or having a reader may take longer than reading regular print. Reading large type may or may not be more time-consuming, depending on the layout of the material and on the nature and severity of the impairment.

MODIFYING TEST SETTING

Tests normally administered in group settings may be administered individually for a variety of purposes. Individual administration may avoid interference with others taking a test in a group. Some disabilities (e.g., attention deficit disorder) make it impractical to test in a group setting. Other alterations may include changing the testing location if it is not wheelchair accessible, providing tables or chairs that provide greater physical support, or altering the lighting conditions for individuals who are visually impaired.

USING ONLY PORTIONS OF A TEST

Another strategy of test accommodation involves the use of portions of a test in assessing persons with disabilities. These procedures are sometimes used in clinical testing when certain subparts of a test require physical, sensory, language, or other capabilities that a test taker with disabilities does not have. This approach is commonly used in cognitive and achievement testing when the physical or sensory limitations of an individual interfere with the ability to perform on a test. For example, if a cognitive ability test includes items presented orally combined with items presented in a written fashion, the orally-presented items might be omitted when the test is given to an individual with a hearing disability as they will not provide an adequate assessment of that individual's cognitive ability. Results on such items are more likely to reflect the individual's hearing difficulty rather than his or her true cognitive ability. Although omitting test items may represent an effective accommodation technique, it may also prevent the test from adequately measuring the intended skills or abilities, especially if those skills or abilities are of central interest. For example, it should be noted that eliminating a portion of the test may not be appropriate in situations such as certification testing and employment testing where the construct measured by the each portion may represent a separate and necessary job or occupational requirement.

USING SUBSTITUTE TESTS OR ALTERNATE ASSESSMENTS

One additional modification is to replace a test standardized on the general population with a test or alternate assessment that has been specially designed for individuals with disabilities. More valid results may be obtained through the use of a test specifically designed for use with individuals with disabilities. Although a substitute test may represent a desirable accommodation solution, it may be difficult to find an adequate replacement that measures the same construct with comparable technical quality,

and for which scores can be placed on the same scale as the original test.

Using Modifications in Different Testing Contexts

There are important contextual differences between the individualized use of tests, as in the case of clinical diagnosis, and group or large-scale testing, as in the case of testing for academic achievement, employment, credentialing, or admissions.

Individual diagnostic testing is conducted typically for clinical or educational purposes. In these contexts a highly qualified test professional (e.g., a licensed or certified psychologist) is responsible for the entire assessment process of test selection, administration, interpretation, and reporting of results. The test professional seeks to gather appropriate information about the client's specific disability and preferred modality of communication and uses this information to determine the accommodations appropriate for the test taker. During the assessment process, any modified tests are used along with other assessment methods to collect data about the client's functioning in relevant areas. Inferences are then made based on this multitude of information. Test modifications may be used during assessment not only out of necessity but also as a source of clinical insight about the client's functioning. For example, a test taker with obsessive compulsive disorder may be allowed to continue to complete a test item, subtest, or a total test beyond the standardized time limits. Although in such cases the performance of the test taker cannot be judged according to the standardized scoring standards, the fact that the test taker could produce a successful performance with extra time often aids clinical interpretation.

The use of test modifications in large-scale testing is different, however. Large-scale testing is used for purposes such as measurement of academic achievement, program evaluation, credentialing, licensure, and employment. In these contexts, a standardized test usually is

PART II / TESTING INDIVIDUALS WITH DISABILITIES

administered to all test participants. Large numbers of test takers are not uncommon, and decisions may in some cases be made solely on the basis of test information, as in the case of a test used as an initial screening device in an employment context. In some cases, decision making requires the comparison of test takers, as in selection or admission contexts where the number of applicants may greatly exceed the number of available openings. This context highlights the need for concern for fairness to all parties, as comparisons must be made between test scores obtained by individuals with disabilities taking modified tests and scores obtained by individuals under regular conditions. While test takers should not be disadvantaged due to a disability not relevant to the construct the test is intended to assess, the resulting accommodation should not put those taking a modified test at an undue advantage over those tested under regular conditions. As research on the comparability of scores under regular and modified conditions is sometimes limited, decisions about appropriate accommodation in these contexts involve important and difficult professional judgments.

Reporting Scores on Modified Tests

The practice of reporting scores on modified tests varies in different contexts. In individual testing, the test professional commonly reports when tests have been administered in a nonstandardized fashion when providing test scores. Typically, the steps used in making test accommodations or modifications are described in the test report, and the validity of the inferences resulting from the modified test scores is discussed. This practice of reporting the nature of modifications is consistent with implied requirements to communicate information as to the nature of the assessment process if the modifications impact the reliability of test scores or the validity of inferences drawn from test scores.

On the other hand, the reporting of test scores from modified tests in large-scale test-

ing has created considerable debate. Often when scores from a nonstandardized version of a test are reported, the score report contains an asterisk next to the score or some other designation, often called a *flag*, to indicate that the test administration was modified. Sometimes recipients of these special designations are informed of the meaning of the designation; many times no information is provided about the nature of the modification made. Some argue that reporting scores from nonstandard test administrations without special identification misleads test users and perhaps even harms test takers with disabilities, whose scores may not accurately reflect their abilities. Others, however, argue that identifying scores of test takers with disabilities as resulting from nonstandard administrations unfairly labels these test takers as persons with disabilities, stigmatizes them, and may deny them the opportunity to compete equally with test takers without disabilities when they might otherwise be able to do so. Federal laws and the laws of most states bar discrimination against persons with disabilities, require individualized reasonable accommodations in testing, and limit practices that could stigmatize persons with disabilities, particularly in educational, admissions, credentialing, and employment testing.

The fundamental principles relevant here are that important information about test score meaning should not be withheld from test users who interpret and act on the test scores, and that irrelevant information should not be provided. When there is sufficient evidence of score comparability across regular and modified administrations, there is no need for any sort of flagging. When such evidence is lacking, an undifferentiated flag provides only very limited information to the test user, and specific information about the nature of the modification is preferable, if permitted by law.

STANDARDS

TESTING INDIVIDUALS WITH DISABILITIES / PART II

Standard 10.1

In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement.

Comment: Chapter 1 (Validity) deals more broadly with the critical requirement that a test score reflects the intended construct. The need to attend to the possibility of construct-irrelevant variance resulting from a test taker's disability is an example of this general principle. In some settings, test users are prohibited from inquiring about a test taker's disability, making the standard contingent on test taker self-report of a disability or a need for accommodation.

Standard 10.2

People who make decisions about accommodations and test modification for individuals with disabilities should be knowledgeable of existing research on the effects of the disabilities in question on test performance. Those who modify tests should also have access to psychometric expertise for so doing.

Comment: In some areas there may be little known about the effects of a particular disability on performance on a particular type of test.

Standard 10.3

Where feasible, tests that have been modified for use with individuals with disabilities should be pilot tested on individuals who have similar disabilities to investigate the appropriateness and feasibility of the modifications.

Comment: Although useful guides for modifying tests are available, they do not provide a universal substitute for trying out a modified test. Even when such tryouts are conducted

on samples inadequate to produce norm data, they are useful for checking the mechanics of the modifications. In many circumstances, however, lack of ready access to individuals with similar disabilities, or an inability to postpone decision making, make this unfeasible.

Standard 10.4

If modifications are made or recommended by test developers for test takers with specific disabilities, the modifications as well as the rationale for the modifications should be described in detail in the test manual and evidence of validity should be provided whenever available. Unless evidence of validity for a given inference has been established for individuals with the specific disabilities, test developers should issue cautionary statements in manuals or supplementary materials regarding confidence in interpretations based on such test scores.

Comment: When test developers and users intend that a modified version of a test should be interpreted as comparable to an unmodified one, evidence of test score comparability should be provided.

Standard 10.5

Technical material and manuals that accompany modified tests should include a careful statement of the steps taken to modify the tests to alert users to changes that are likely to alter the validity of inferences drawn from the test score.

Comment: If empirical evidence of the nature and effects of changes resulting from modifying standard tests is lacking, it is impossible to assess the impact of significant modifications. Documentation of the procedures used to modify tests will not only aid in the administration and interpretation of the given test but will also inform others who are modifying tests for people with spe-

PART II / TESTING INDIVIDUALS WITH DISABILITIES

STANDARDS

cific disabilities. This standard should apply to both test developers and test users.

Standard 10.6

If a test developer recommends specific time limits for people with disabilities, empirical procedures should be used, whenever possible, to establish time limits for modified forms of timed tests rather than simply allowing test takers with disabilities a multiple of the standard time. When possible, fatigue should be investigated as a potentially important factor when time limits are extended.

Comment: Such empirical evidence is likely only in the limited settings where a sufficient number of individuals with similar disabilities are tested. Not all individuals with the same disability, however, necessarily require the same accommodation. In most cases, professional judgment based on available evidence regarding the appropriate time limits given the nature of an individual's disability will be the basis for decisions. Legal requirements may be relevant to any decision on absolute time limits.

Standard 10.7

When sample sizes permit, the validity of inferences made from test scores and the reliability of scores on tests administered to individuals with various disabilities should be investigated and reported by the agency or publisher that makes the modification. Such investigations should examine the effects of modifications made for people with various disabilities on resulting scores, as well as the effects of administering standard unmodified tests to them.

Comment: In addition to modifying tests and test administration procedures for people who have disabilities, evidence of validity for inferences drawn from these tests is needed. Validation is the only way to amass knowledge about the usefulness of modified tests

for people with disabilities. The costs of obtaining validity evidence should be considered in light of the consequences of not having usable information regarding the meanings of scores for people with disabilities. This standard is feasible in the limited circumstances where a sufficient number of individuals with the same level or degree of a given disability is available.

Standard 10.8

Those responsible for decisions about test use with potential test takers who may need or may request specific accommodations should (a) possess the information necessary to make an appropriate selection of measures, (b) have current information regarding the availability of modified forms of the test in question, (c) inform individuals, when appropriate, about the existence of modified forms, and (d) make these forms available to test takers when appropriate and feasible.

Standard 10.9

When relying on norms as a basis for score interpretation in assessing individuals with disabilities, the norm group used depends upon the purpose of testing. Regular norms are appropriate when the purpose involves the test taker's functioning relative to the general population. If available, normative data from the population of individuals with the same level or degree of disability should be used when the test taker's functioning relative to individuals with similar disabilities is at issue.

Standard 10.10

Any test modifications adopted should be appropriate for the individual test taker, while maintaining all feasible standardized features. A test professional needs to consider reasonably available information about each test taker's experiences, characteristics,

STANDARDS

TESTING INDIVIDUALS WITH DISABILITIES / PART II

and capabilities that might impact test performance, and document the grounds for the modification.

Standard 10.11

When there is credible evidence of score comparability across regular and modified administrations, no flag should be attached to a score. When such evidence is lacking, specific information about the nature of the modification should be provided, if permitted by law, to assist test users properly to interpret and act on test scores.

Comment: The inclusion of a flag on a test score where an accommodation for a disability was provided may conflict with legal and social policy goals promoting fairness in the treatment of individuals with disabilities. If a score from a modified administration is comparable to a score from a nonmodified administration, there is no need for a flag. Similarly, if a modification is provided for which there is no reasonable basis for believing that the modification would affect score comparability, there is no need for a flag. Further, reporting practices that use asterisks or other nonspecific symbols to indicate that a test's administration has been modified provide little useful information to test users. When permitted by law, if a non-standardized administration is to be reported because evidence does not exist to support score comparability, then this report should avoid referencing the existence or nature of the test taker's disability and should instead report only the nature of the accommodation provided, such as extended time for testing, the use of a reader, or the use of a tape recorder.

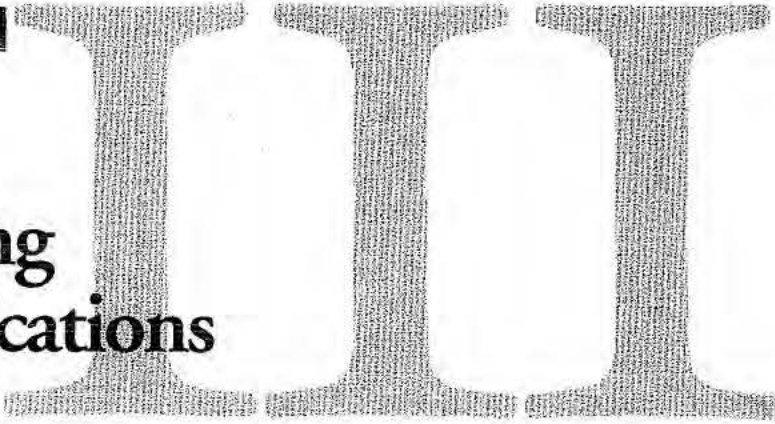
Standard 10.12

In testing individuals with disabilities for diagnostic and intervention purposes, the test should not be used as the sole indicator of the test taker's functioning. Instead, multiple sources of information should be used.

Comment: For example, when assessing the intellectual functioning of persons with mental retardation, results from an individually administered intelligence test are generally supplemented with other pertinent information, such as case history, information about school functioning, and results from other cognitive tests and adaptive behavior measures. In addition, at times a multidisciplinary evaluation (e.g., physical, psychological, linguistic, neurological, etc.) may be needed to yield an accurate picture of the person's functioning.

PART III

Testing Applications

A large, stylized Roman numeral 'III' is rendered in a dense, halftone-like dot pattern. It is positioned to the right of the text 'PART III' and 'Testing Applications'.

11. THE RESPONSIBILITIES OF TEST USERS

Background

Previous chapters have dealt primarily with the responsibilities of those who develop, market, evaluate, or mandate the administration of tests and the rights and obligations of test takers. Many of the standards in these chapters, and in the chapters that follow, refer to the development of tests and their use in specific settings. The present chapter includes standards of a more general nature that apply in almost all measurement contexts. In particular, attention is centered on the responsibilities of those who may be considered the *users* of tests. This group includes psychologists, educators, and other professionals who select the specific instruments or supervise test administration—on their own authority or at the behest of others. It also includes all individuals who actively participate in the interpretation and use of test results, other than the test takers themselves.

It is presumed that a legitimate educational, psychological, or employment purpose justifies the time and expense of test administration. In most settings, the user communicates this purpose to those who have a legitimate interest in the measurement process and subsequently conveys the implications of examinee performance to those entitled to receive the information. Depending on the measurement setting, this group may include individual test takers, parents and guardians, educators, employers, policymakers, the courts, or the general public.

Where administration of tests or use of test data is mandated for a specific population by governmental authorities, educational institutions, licensing boards, or employers, the developer and user of an instrument may be essentially the same. In such settings, there often is no clear separation between the professional responsibilities of those who produce the instrument and those who administer the test and interpret the results. Instruments pro-

duced by independent publishers, on the other hand, present a somewhat different picture. Typically, these tests will be used with a variety of populations and for diverse purposes.

The conscientious developer of a standardized test attempts to screen and educate potential users. Furthermore, most publishers and test sponsors work vigorously to prevent the misuse of standardized measures and the misinterpretation of individual scores and group averages. Test manuals often illustrate sound and unsound interpretations and applications. Some identify specific practices that are not appropriate and should be discouraged. Despite the best efforts of test developers, however, appropriate test use and sound interpretation of test scores are likely to remain primarily the responsibility of the test user.

Test takers, parents and guardians, legislators, policymakers, the media, the courts, and the public at large often yearn for unambiguous interpretations of test data. In particular, they often tend to attribute positive or negative results, including group differences, to a single factor or to the conditions that prevail in one social institution—most often, the home or the school. These consumers of test data frequently press for explicit rationales for decisions that are based only in part on test scores. The wise test user helps all interested parties understand that sound decisions regarding test use and score interpretation involve an element of professional judgment. It is not always obvious to the consumers that the choice of various information-gathering procedures often involves experience that is not easily quantified or verbalized. The user can help them appreciate the fact that the weighting of quantitative data, educational and occupational information, behavioral observations, anecdotal reports, and other relevant data often cannot be specified precisely.

THE RESPONSIBILITIES OF TEST USERS / PART III

Because of the appearance of objectivity and numerical precision, test data are sometimes allowed to totally override other sources of evidence about test takers. There are circumstances in which selection based exclusively on test scores may be appropriate. For example, this may be the case in pre-employment screening. But in educational and psychological settings, test users are well advised, and may be legally required, to consider other relevant sources of information on test takers, not just test scores. In the latter situations, the psychologist or educator familiar with the local setting and with local test takers is best qualified to integrate this diverse information effectively.

As reliance on test results has grown in recent years, greater pressure has been placed on test users to explain to the public the rationale for test-based decisions. More than ever before, test users are called upon to defend their testing practices. They do this by documenting that their test uses and score interpretations are supported by measurement authorities for the given purpose, that the inferences drawn from their instruments are validated for use with a given population, and that the results are being used in conjunction with other information, not in isolation. If these conditions are met, the test user can convincingly defend the decisions made or the administrative actions taken in which tests played a part.

It is not appropriate for these *Standards* to dictate minimal levels of test-criterion correlation, classification accuracy, or reliability for any given purpose. Such levels depend on whether decisions must be made immediately on the strength of the best available evidence, however weak, or whether decisions can be delayed until better evidence becomes available. But it is appropriate to expect the user to ascertain what the alternatives are, what the quality and consequences of these alternatives are, and whether a delay in decision making would be beneficial. Cost-benefit compromises become necessary in test use, as they often are in test development. It should be noted, how-

ever, that in some contexts legal requirements may place limits on the extent to which such compromises can be made. As with standards for the various phases of test development, when relevant standards are not met in test use, the reasons should be persuasive. The greater the potential impact on test takers, for good or ill, the greater the need to identify and satisfy the relevant standards.

In selecting a test and interpreting a test score, the test user is expected to have a clear understanding of the purposes of the testing and its probable consequences. The knowledgeable user has definite ideas on how to achieve these purposes and how to avoid bias, unfairness, and undesirable consequences. In subscribing to these *Standards*, test publishers and agencies mandating test use agree to provide information on the strengths and weaknesses of their instruments. They accept the responsibility to warn against likely misinterpretations by unsophisticated interpreters of individual scores or aggregated data. However, the ultimate responsibility for appropriate test use and interpretation lies predominantly with the test user. In assuming this responsibility, the user must become knowledgeable about a test's appropriate uses and the populations for which it is suitable. The user must also become adept, particularly in statewide and community-wide assessment programs, in communicating the implications of test results to those entitled to receive them.

In some instances, users may be obligated to collect additional evidence about a test's technical quality. For example, if performance assessments are locally scored, evidence of the degree of inter-scorer agreement may be required. Users also should be alert to the probable local consequences of test use, particularly in the case of large-scale testing programs. If the same test material is used in successive years, users should actively monitor the program to ensure that reuse has not compromised the integrity of the results.

Some of the standards that follow reiterate ideas contained in other chapters, principally chapter 5 “Test Administration, Scoring, and Reporting,” chapter 7 “Fairness in Testing and Test Use,” chapter 8 “Rights and Responsibilities of Test Takers,” and chapter 13 “Educational Testing and Assessment.” This repetition is intentional. It permits an enumeration in one chapter of the major obligations that must be assumed largely by the test administrator and user, though these responsibilities may refer to topics that are covered more fully in other chapters.

Standard 11.1

Prior to the adoption and use of a published test, the test user should study and evaluate the materials provided by the test developer. Of particular importance are those that summarize the test’s purposes, specify the procedures for test administration, define the intended populations of test takers, and discuss the score interpretations for which validity and reliability data are available.

Comment: A prerequisite to sound test use is knowledge of the materials accompanying the instrument. As a minimum, these include manuals provided by the test developer. Ideally, the user should be conversant with relevant studies reported in the professional literature. The degree of reliability and validity required for sound score interpretations depends on the test’s role in the assessment process and the potential impact of the process on the people involved. The test user should be aware of legal restrictions that may constrain the use of the test. On occasion, professional judgment may lead to the use of instruments for which there is little documentation of validity for the intended purpose. In these situations, the user should interpret scores cautiously and take care not to imply that the decisions or inferences are based on test results that are well-documented with respect to reliability or validity.

Standard 11.2

When a test is to be used for a purpose for which little or no documentation is available, the user is responsible for obtaining evidence of the test’s validity and reliability for this purpose.

Comment: The individual who uses test scores for purposes that are not specifically recommended by the test developer is responsible for collecting the necessary validity evidence. Support for such uses may sometimes be found in the professional literature. If previous evidence is not sufficient, then additional data should be

STANDARDS

THE RESPONSIBILITIES OF TEST USERS / PART III

collected. The provisions of this standard should not be construed to prohibit the generation of hypotheses from test data. For example, though some clinical tests have limited or contradictory validity evidence for common uses, clinicians generate hypotheses based appropriately on examinee responses to such tests. However, these hypotheses should be clearly labeled as tentative. Interested parties should be made aware of the potential limitations of the test scores in such situations.

Standard 11.3

Responsibility for test use should be assumed by or delegated only to those individuals who have the training, professional credentials, and experience necessary to handle this responsibility. Any special qualifications for test administration or interpretation specified in the test manual should be met.

Comment: Test users should not attempt to interpret the scores of test takers whose special needs or characteristics are outside the range of the user's qualifications. This standard has special significance in areas such as clinical testing, forensic testing, testing in special education, testing people with disabilities or limited exposure to the dominant culture, and in other such situations where potential impact is great. When the situation falls outside the user's experience, assistance should be obtained. A number of professional organizations have codes of ethics that specify the qualifications of those who administer tests and interpret scores.

Standard 11.4

The test user should have a clear rationale for the intended uses of a test or evaluation procedure in terms of its validity and contribution to the assessment and decision-making process.

Comment: Justification for the role of each instrument in selection, diagnosis, classification, and decision making should be arrived

at before test administration, not afterwards. Preferably, the rationale should be available in printed materials prepared by the test publisher or by the user.

Standard 11.5

Those who have a legitimate interest in an assessment should be informed about the purposes of testing, how tests will be administered, the factors considered in scoring examinee responses, how the scores are typically used, how long the records will be retained, and to whom and under what conditions the records may be released.

Comment: This standard has greater relevance and application to educational and clinical testing than to employment testing. In most uses of tests for screening job applicants and applicants to educational programs, for licensing professionals and awarding credentials, or for measuring achievement, the purposes of testing and the uses to be made of the test scores are obvious to the examinee. Nevertheless, it is wise to communicate this information at least briefly even in these settings. In some situations, however, the rationale for the testing may be clear to relatively few test takers. In such settings, a more detailed and explicit discussion may be called for. Retention and release of records, even when such release would clearly benefit the examinee, are often governed by statutes or institutional practices. As relevant, examinees should be informed about these constraints and procedures.

Standard 11.6

Unless the circumstances clearly require that the test results be withheld, the test user is obligated to provide a timely report of the results that is understandable to the test taker and others entitled to receive this information.

Comment: The nature of score reports is often dictated by practical considerations. In some

cases only a terse printed report may be feasible. In others, it may be desirable to provide both an oral and a written report. The interpretation should vary according to the level of sophistication of the recipient. When the examinee is a young child, an explanation of the test results is typically provided to parents or guardians. Feedback in the form of a score report or interpretation is not typically provided when tests are administered for personnel selection or promotion.

Standard 11.7

Test users have the responsibility to protect the security of tests, to the extent that developers enjoin users to do so.

Comment: When tests are used for purposes of selection, licensure, or educational accountability, the need for rigorous protection of test security is obvious. On the other hand, when educational tests are not part of a high-stakes program, some publishers consider teacher review of test materials to be a legitimate tool in clarifying teacher perceptions of the skills measured by a test. Consistency and clarity in the definition of acceptable and unacceptable practices is critical in such situations. When tests are involved in litigation, inspection of the instruments should be restricted—to the extent permitted by law—to those who are legally or ethically obligated to safeguard test security.

Standard 11.8

Test users have the responsibility to respect test copyrights.

Comment: Legally and ethically, test users may not reproduce copyrighted materials for routine test use without consent of the copyright holder. These materials—in both paper and electronic form—include test items, ancillary forms such as answer sheets or profile forms, scoring templates, conversion tables of raw scores to derived scores, and tables of norms.

Standard 11.9

Test users should remind test takers and others who have access to test materials that the legal rights of test publishers, including copyrights, and the legal obligations of other participants in the testing process may prohibit the disclosure of test items without specific authorization.

Standard 11.10

Test users should be alert to the possibility of scoring errors; they should arrange for rescoring if individual scores or aggregated data suggest the need for it.

Comment: The costs of scoring error are great, particularly in high-stakes testing programs. In some cases, rescoring may be requested by the test taker. If such a test taker right is recognized in published materials, it should be respected. In educational testing programs, users should not depend entirely on test takers to alert them to the possibility of scoring errors. Monitoring scoring accuracy should be a routine responsibility of testing program administrators wherever feasible.

Standard 11.11

If the integrity of a test taker's scores is challenged, local authorities, the test developer, or the test sponsor should inform the test takers of their relevant rights, including the possibility of appeal and representation by counsel.

Comment: Proctors in entrance or licensure testing programs may report irregularities in the test process that result in challenges. University admissions officers may raise challenges when test scores are grossly inconsistent with other applicant information. Test takers should be apprised of their rights in such situations.

STANDARDS

THE RESPONSIBILITIES OF TEST USERS / PART III

Standard 11.12

Test users or the sponsoring agency should explain to test takers their opportunities, if any, to retake an examination; users should also indicate whether the earlier as well as later scores will be reported to those entitled to receive the score reports.

Comment: Some testing programs permit test takers to retake an examination several times, to cancel scores, or to have scores withheld from potential recipients. If test takers have such privileges, they and score recipients should be so informed.

Standard 11.13

When test-taking strategies that are unrelated to the domain being measured are found to enhance or adversely affect test performance significantly, these strategies and their implications should be explained to all test takers before the test is administered. This may be done either in an information booklet or, if the explanation can be made briefly, along with the test directions.

Comment: Test-taking strategies, such as guessing, skipping time-consuming items, or initially skipping and then returning to difficult items as time allows, can influence test scores positively or negatively. The effects of various strategies depend on the scoring system used and aspects of item and test design such as speededness or the number of response alternatives provided in multiple-choice items. Differential use of such strategies by test takers can affect the validity and reliability of test score interpretations. The goal of test directions should be to convey information on the possible effectiveness of various strategies and, thus, to provide all test takers an equal opportunity to perform optimally. The use of such strategies by all test takers should be encouraged if their effect facilitates performance and discouraged if their effect interferes with performance.

116

Standard 11.14

Test users are obligated to protect the privacy of examinees and institutions that are involved in a measurement program, unless a disclosure of private information is agreed upon, or is specifically authorized by law.

Comment: Protection of the privacy of individual examinees is a well-established principle in psychological and educational measurement. In some instances, test takers and test administrators may formally agree to a lesser degree of protection than the law appears to require. In other circumstances, test users and testing agencies may adopt more stringent restrictions on the communication and sharing of test results than relevant law dictates. The more rigorous standards sometimes arise through the codes of ethics adopted by relevant professional organizations. In some testing programs the conditions for disclosure are stated to the examinee prior to testing, and taking the test can constitute agreement for the disclosure of test score information as specified. In other programs, the test taker or his/her parents or guardians must formally agree to any disclosure of test information to individuals or agencies other than those specified in the test administrator's published literature. It should be noted that the right of the public and the media to examine the aggregate test results of public school systems is guaranteed in some states.

Standard 11.15

Test users should be alert to potential misinterpretations of test scores and to possible unintended consequences of test use; users should take steps to minimize or avoid foreseeable misinterpretations and unintended negative consequences.

Comment: Well-meaning, but unsophisticated, audiences may adopt simplistic interpretations of test results or may attribute high or low scores or averages to a single causal factor.

PART III / THE RESPONSIBILITIES OF TEST USERS

STANDARDS

Experienced test users can sometimes anticipate such misinterpretations and should try to prevent them. Obviously, not every unintended consequence can be anticipated. What is required is a reasonable effort to prevent negative consequences and to encourage sound interpretations.

Standard 11.16

Test users should verify periodically that their interpretations of test data continue to be appropriate, given any significant changes in their population of test takers, their modes of test administration, and their purposes in testing.

Comment: Over time, a gradual change in the demographic characteristics of an examinee population may significantly affect the inferences drawn from group averages. The accommodations made in test administration in recognition of examinee disabilities or in response to unforeseen circumstances may also affect interpretations.

Standard 11.17

In situations where the public is entitled to receive a summary of test results, test users should formulate a policy regarding timely release of the results and apply that policy consistently over time.

Comment: In school testing programs, districts commonly viewed as a coherent group may avoid controversy by adopting the same policies regarding the release of test results. If one district routinely releases aggregated data in much greater detail than another, groundless suspicions can develop that information is being suppressed in the latter district.

Standard 11.18

When test results are released to the public or to policymakers, those responsible for the release should provide and explain any

supplemental information that will minimize possible misinterpretations of the data.

Comment: Preliminary briefings prior to the release of test results can give reporters for the news media an opportunity to assimilate relevant data. Misinterpretation can often be the result of the limited time reporters have to prepare media reports or inadequate presentation of information that bears on test score interpretation. It should be recognized, however, that the interests of the media are not always consistent with the intended purposes of measurement programs.

Standard 11.19

When a test user contemplates an approved change in test format, mode of administration, instructions, or the language used in administering the test, the user should have a sound rationale for concluding that validity, reliability, and appropriateness of norms will not be compromised.

Comment: In some instances, minor changes in format or mode of administration may be reasonably expected, without evidence, to have little or no effect on validity, reliability, and appropriateness of norms. In other instances, however, changes in format or administrative procedures can be assumed a priori to have significant effects. When a given modification becomes widespread, consideration should be given to validation and norming under the modified conditions.

Standard 11.20

In educational, clinical, and counseling settings, a test taker's score should not be interpreted in isolation; collateral information that may lead to alternative explanations for the examinee's test performance should be considered.

Comment: It is neither necessary nor feasible to make an intensive review of every test taker's

STANDARDS

THE RESPONSIBILITIES OF TEST USERS / PART III

score. In some settings there may be little or no collateral information of value. In counseling, clinical, and educational settings, however, considerable relevant information is likely to be available. Obvious alternative explanations of low scores include low motivation, limited fluency in the language of the test, unfamiliarity with cultural concepts on which test items are based, and perceptual or motor impairments. In clinical and counseling settings, the test user should not ignore how well the test taker is functioning in daily life.

Standard 11.21

Test users should not rely on computer-generated interpretations of test results unless they have the expertise to consider the appropriateness of these interpretations in individual cases.

Comment: The scoring agency has the responsibility of documenting the basis for the interpretations. The user of a computerized scoring and reporting service has the obligation to be familiar with the principles on which such interpretations were derived. The user should have the ability to evaluate a computer-based score interpretation in the light of other relevant evidence on each test taker. Automated, narrative reports are not a substitute for sound professional judgment.

Standard 11.22

When circumstances require that a test be administered in the same language to all examinees in a linguistically diverse population, the test user should investigate the validity of the score interpretations for test takers believed to have limited proficiency in the language of the test.

Comment: The achievement, abilities, and traits of examinees who do not speak the language of the test as their primary language may be seriously mismeasured by the test.

The scores of test takers with severe linguistic limitations will probably be meaningless. If language proficiency is not relevant to the purposes of testing, the test user should consider excusing these individuals, without prejudice, from taking the test and substituting alternative evaluation methods. However, it is recognized that such actions may be impractical, unnecessary, or legally unacceptable in some settings.

Standard 11.23

If a test is mandated for persons of a given age or all students in a particular grade, users should identify individuals whose disabilities or linguistic background indicates the need for special accommodations in test administration and ensure that these accommodations are employed.

Comment: Appropriate accommodations depend upon the nature of the test and the needs of the test taker. The mandating authority has primary responsibility for defining the acceptable accommodations for various categories of test takers. The user must take responsibility for identifying those test takers who fall within these categories and implement the appropriate accommodations.

Standard 11.24

When a major purpose of testing is to describe the status of a local, regional, or particular examinee population, the program criteria for inclusion or exclusion of individuals should be strictly adhered to.

Comment: In census-type programs, biased results can arise from the exclusion of particular subgroups of students. Financial and other advantages may accrue either from exaggerating or from reducing the proportion of high-achieving or low-achieving students. Clearly, these are unprofessional practices.

12. PSYCHOLOGICAL TESTING AND ASSESSMENT

Background

This chapter addresses issues important to professionals who use psychological tests with their clients. Topics include test selection and administration, test interpretation, collateral information used in psychological testing, types of tests, and purposes of testing. The types of psychological tests reviewed in this chapter include cognitive and neuropsychological; adaptive, social, and problem behavior; family and couples; personality; and vocational. In addition, the chapter includes an overview of four common uses of psychological tests: diagnosis; intervention planning and outcome evaluation; legal and governmental decisions; and personal awareness, growth, and action.

Employment testing is another context in which psychological testing is used. The standards in this chapter are applicable to those employment settings in which individual in-depth assessment is conducted (e.g., an evaluation of a candidate for a senior executive position). Employment settings in which tests are designed to measure specific job-related characteristics across multiple candidates are treated in the text and standards of chapter 14.

For all professionals who use tests, knowledge of cultural background and physical capabilities that influence (a) a test taker's development, (b) the methods for obtaining and conveying information, and (c) the planning and implementation of interventions is critical. Therefore, readers are encouraged to review chapters 7, 8, 9, and 10 that discuss fairness and bias in testing, the rights and responsibilities of test takers, testing individuals of diverse linguistic backgrounds, and testing individuals with disabilities. Readers will find important additional detail on validity; reliability; test development; scaling; test administration, scoring, and reporting; and general responsibilities of test users in chapters 1, 2, 3, 4, 5, and 11, respectively.

The use of tests provides one method of collecting information within the larger framework of a psychological assessment of an individual. Typically, psychological assessments involve an interaction between a professional who is trained and experienced in testing and a client. Clients may include patients, counselees, parents, employees, employers, attorneys, students, and other responsible parties who are test takers or who use the test results contained in psychological reports.

The results from tests and inventories, used within the context of a psychological assessment, may help the professional to understand the client more fully and to develop more informed and accurate hypotheses, inferences, and decisions about a client's situation. A psychological assessment is a comprehensive examination undertaken to answer specific questions about a client's psychological functioning during a particular time interval or to predict a client's psychological functioning in the future. An assessment may include administering and scoring tests, and interpreting test scores, all within the context of the individual's personal history. Inasmuch as test scores characteristically are interpreted in the context of other information about the client, an individual psychological assessment usually also includes interviewing the client; observing client behavior; reviewing educational, psychological, and other relevant records; and integrating these findings with other information that may be provided by third parties. The tasks of a psychological assessment—collecting, evaluating, integrating, and reporting salient information relevant to those aspects of a client's functioning that are under examination—comprise a complex and sophisticated set of professional activities.

The interpretation of tests and inventories can be a valuable part of the intervention process and, if used appropriately, can provide useful information to clients as well as to other users

of the test interpretation. For example, the results of tests and inventories may be used to assess the psychological functioning of an individual; to assign diagnostic classifications; to detect neuropsychological impairment; to assess cognitive and personality strengths, vocational interests, and values; to determine developmental stages; and to evaluate treatment outcomes. Test results also may provide information used to make decisions that have a powerful and lasting impact on people's lives (e.g., vocational and educational decision making; diagnosis; treatment planning; selection decisions; intervention and outcome evaluation; parole, sentencing, civil commitment, child custody, and competency to stand trial decisions; and personal injury litigation).

TEST SELECTION AND ADMINISTRATION

Prior to beginning the assessment process, the test taker should understand who will have access to the test results and the written report, how test results will be shared with the test taker, and if and when decisions based on the test results will be shared with the test taker and/or a third party. The assessment process begins by clarifying, as much as is possible, the reasons for which a client is presented for assessment. Guided by these reasons or other relevant concerns, the tests, inventories, and diagnostic procedures to be used are chosen, and other sources of information needed to evaluate the client and the referral issues are identified. The professional reviews more than the name of the test in choosing a test and is guided by the validity and reliability evidence and the applicability of the normative data available in the test's accumulated research literature. In addition to being thoroughly versed in proper administrative procedure, the professional is responsible for being familiar with the validity and reliability evidence for the intended use and purposes of the tests and inventories selected and for being prepared to develop a logical analysis that supports the various facets of the assessment and the inferences made from the assessment.

Validity and reliability considerations are paramount, but the demographic characteristics (e.g., gender, age, income, sociocultural and language background, education and other socioeconomic variables) of the group for which the test was originally constructed and for which initial and subsequent normative data are available also are important test selection issues. Selecting a test with demographically appropriate normative groups relevant for the client being tested is important to the generalizability of the inferences that the professional seeks to make. Sometimes the items or tasks contained in a test are designed for a particular group and are viewed as irrelevant for another group. A test constructed for one group may be applied to other groups with appropriate qualifications that explain the test choice based on the supporting research data and on professional experience.

The selection of psychological tests and inventories, for a particular client, often is individualized. However, in some settings a predetermined battery of tests may be taken by all participants, and group interpretations may be provided. The test taker may be a child, an adolescent, or an adult. The settings in which the tests or inventories are used include (but are not limited to) preschool, elementary, middle, or secondary schools; colleges or universities; pre-employment or employment settings; mental health or outpatient clinics; hospitals; prisons; or professionals' offices.

Professionals who oversee testing and assessment are responsible for ensuring that all persons who administer and score tests have received the appropriate education and training needed to perform these tasks. In addition, they are responsible in group testing situations for ensuring that the individuals who use the test results are trained to interpret the scores properly.

When conducting psychological testing, standardized test administration procedures should be followed. When nonstandard administration procedures are needed, they are to be described and justified. Professionals

PART III / PSYCHOLOGICAL TESTING AND ASSESSMENT

also are responsible for ensuring that testing conditions are appropriate. For example, the examiner may need to determine if the client is capable of reading at the level required, and if clients with vision, hearing, or neurological disabilities are adequately accommodated. Finally, professionals are responsible for protecting the confidentiality and security of the test results and the testing materials.

One advantage of individually administered measures is the opportunity to observe and adjust testing conditions as needed. In some circumstances, test administration may provide the opportunity for skilled examiners to carefully observe the performance of persons under standardized conditions. For example, their observations may allow them to more accurately record behaviors being assessed, to understand better the manner in which persons arrive at their answers, to identify personal strengths and weaknesses, and to make modifications in the testing process. Thus, the observations of trained professionals can be important to all aspects of test use.

TEST SCORE INTERPRETATION

Test scores ideally are interpreted in light of the available normative data, the psychometric properties of the test, the temporal stability of the constructs being measured, and the effect of moderator variables and demographic characteristics (e.g., gender, age, income, sexual orientation, sociocultural and language background, education, and other socioeconomic variables) on test results. The professional rarely has the resources available to personally conduct the research or to assemble representative norms needed to make accurate inferences about each individual client's current and future functioning. Therefore, the professional may rely on the research and the body of scientific knowledge available for the test that warrants appropriate inferences. Presentation and analyses of validity and reliability evidence often are not needed in a written report, but the professional

strives to understand, and prepares to articulate, such evidence as the need arises.

Tests and inventories that meet high technical standards of quality are a necessary but not a sufficient condition to ensure the responsible use and interpretation of test scores. The level of competence of the professional who interprets the scores and integrates the inferences derived from psychological tests depends upon the educational and experiential qualifications of the professional. With experience, professionals learn that the challenges in psychological test score interpretation increase in magnitude along a continuum of professional judgment with brief screening inventories at one end of the continuum and comprehensive multidimensional assessments at the other. For example, the interpretations of achievement and ability test scores, personality test scores, and batteries of neuropsychological test scores represent points on a continuum that require increasing levels of specialized knowledge, judgment, and skill by an experienced professional regardless of the soundness of the technical characteristics of the tests being used. The education and experience necessary to administer group tests and/or proctor computer-administered tests generally are less stringent than are the qualifications necessary to interpret individually administered tests. The use and interpretation of individually administered tests requires completion of rigorous educational and applied training, a high degree of professional judgment, appropriate credentialing, and adherence to the professional's ethical guidelines.

When making inferences about a client's past, present, and future behaviors and other characteristics from test scores, the professional reviews the literature to develop familiarity with supporting evidence. When there is strong evidence supporting the reliability and validity of a test, including its applicability to the client being assessed, the professional's ability to draw inferences increases. Nevertheless, the professional still corroborates results from testing with additional information from a variety of sources

such as interviews and results from other tests. When an inference is based on a single study or based on several studies whose samples are not representative of the client, the professional is more cautious about the inferences. Corroborating data from the assessment's multiple sources of information—including stylistic and test-taking behaviors inferred from observations during the test—will strengthen the confidence placed in the inference. Importantly, data that are not supportive of the inference are acknowledged and either reconciled or noted as limits to the confidence placed in the inference.

An interpretation of a test taker's test scores based upon existing research examines not only the demonstrated relationship between the scores and the criterion or criteria, but also the appropriateness of the latter. The criterion and the chosen predictor test or tests are subjected to a similar examination to understand the degree to which their underlying constructs are congruent with the inferences under consideration.

Threats to the interpretability of obtained scores are minimized by clearly defining how particular psychological tests are used. These threats occur as a result of construct-irrelevant variance (i.e., aspects of the test that are not relevant to the purpose of the test scores) and construct underrepresentation (i.e., important facets relevant to the purpose of the testing, but for which the test does not account). A client's response bias is another example of a construct-irrelevant component that may significantly skew the obtained scores, possibly rendering the scores uninterpretable. In situations where response bias is anticipated, the professional may choose a test that has scales (e.g., faking good, faking bad, social desirability, percent yes, percent no) that clarify the threats to validity from the test taker's response bias. In so doing, the professional may be able to assess the degree to which test takers are acquiescing to the perceived demands of the test administrator or attempting to portray themselves as impaired by "faking bad," or well-functioning by "faking good." In interpreting the test taker's obtained

response bias score(s), the evidence of validity for constructs underlying each response bias scale, each scale's internal consistency, its interrelations with other scales, and evidence of validity are considered.

For some purposes, including career counseling and neuropsychological assessment, test batteries frequently are used. Such batteries often include tests of verbal ability, numerical ability, nonverbal reasoning, mechanical reasoning, clerical speed and accuracy, spatial ability, and language usage. Some batteries also include interest and personality inventories. When psychological test batteries incorporate multiple methods and scores, patterns of test results frequently are interpreted to reflect a construct or even an interaction among constructs underlying test performances. Higher order interactions among the constructs underlying configurations of test outcomes may be postulated on the basis of test score patterns. The literature reporting evidence of reliability and validity that supports the proposed interpretations should be identifiable. If the literature is incomplete, the resulting inferences may be presented with the qualification that they are hypotheses for future verification rather than probabilistic statements that imply some known validity evidence.

COLLATERAL INFORMATION USED IN PSYCHOLOGICAL TESTING AND PSYCHOLOGICAL ASSESSMENT

The quality of psychological testing and psychological assessment is enhanced by obtaining credible collateral information from various third-party sources such as teachers, personal physicians, family members, and school or employment records. Psychological testing also is enhanced by using various methods to acquire information. Structured behavioral observations, checklists and ratings, interviews, and criterion- and norm-referenced measures are but a few of the methods that may be used to acquire information. The use of psychological tests also can be enhanced by acquiring information about multiple traits or attributes to help characterize a person. For example, an

PART III / PSYCHOLOGICAL TESTING AND ASSESSMENT

evaluation of career goals may be enhanced by obtaining a history of current and prior employment as well as by administering tests to assess academic aptitude and achievement, vocational interests, work values, and personality and temperament characteristics. The availability of information on multiple traits or attributes, when acquired from various sources and through the use of various methods, enables professionals to assess more accurately an individual's psychosocial functioning and facilitates more effective decision making.

Types of Psychological Tests

For purposes of this chapter, the types of psychological tests have been divided into five categories: cognitive and neuropsychological tests; adaptive, social, and problem behavior tests; family and couples tests; personality tests; and vocational tests.

COGNITIVE AND NEUROPSYCHOLOGICAL TESTING

Tests often are used to assess various classes of cognitive and neuropsychological functioning including intelligence; broad ability domains (e.g., verbal, quantitative, and spatial abilities); and more focused domains (e.g., attention, sensorimotor functions, perception, learning, memory, reasoning, executive functions, and language). Overlap may occur in the constructs that are assessed by tests of differing functions or domains. In common with other types of tests, cognitive and neuropsychological tests require a minimally sufficient level of test-taker attentional capacity.

Cognitive Ability. Measures designed to quantify cognitive abilities are among the most widely administered tests. The interpretation of cognitive ability tests is guided by the theoretical constructs used to develop the test.

Many cognitive ability tests consist of multidimensional test batteries that are designed to assess a broad range of abilities and skills. Individually administered test batteries also are required for testing for purposes such as diag-

nosing a cognitive disorder. Test results are used to draw inferences about a person's overall level of intellectual functioning as well as strengths and weaknesses in various cognitive abilities. Because each test in a battery examines a different function, ability, skill, or combination thereof, the test taker's performance can be understood best when scores are not combined or aggregated, but rather when each score is interpreted within the context of all other scores and other assessment data. For example, low scores on timed tests alert the examiner to slowed responding as a problem that may not be apparent if scores on different kinds of tests are combined.

Attention. Attention refers to that class of functioning that encompasses arousal, establishment and deployment of sets, sustained attention, and vigilance as constructs. Tests may measure levels of alertness, orientation, and localization; the ability to focus, shift, and maintain attention and to track one or more stimuli under various conditions; span of attention; information processing speed and choice reaction time; and short-term information storage capacity. Scores for each aspect of attention that has been examined should be reported individually so that the nature of an attention disorder can be clarified.

Motor, Sensorimotor Functions, and Lateral Preferences. Visual, auditory, somatosensory and other sensory sensitivity and discrimination can be measured by simple motor or verbal responses to selective stimulation upon command.

Perception and Perceptual Organization/Integration. This class of functioning involves reasoning and judgment as they relate to the processing and elaboration of complex sensory combinations and inputs. Tests of perception may emphasize immediate perceptual processing but also may require conceptualizations that involve some reasoning and judgmental processes. Some tests have a motor component ranging from a simple motor response to an elaborate construction. Also,

some of these tests penalize the test taker for slow performance that may be caused by something other than perceptual dysfunction.

Learning and Memory. This class of functions involves the acquisition and retention of information beyond the attentional requirements of immediate or short-term information processing and storage. These tests may measure acquisition of new information through various sensory channels and by means of assorted test formats (e.g., word lists, prose passages, geometric figures, formboards, digits, and musical melodies). Memory tests also may require retention and recall of old information (e.g., personal data as well as commonly learned facts and skills).

Abstract Reasoning and Categorical Thinking. Tests of reasoning and thinking vary widely. They assess the examinee's ability to infer relationships or to respond to changing environmental circumstances and to act in goal-oriented situations.

Executive Functions. This class of functions is involved in the organized performances that are necessary for the independent, purposive and effective attainment of personal goals in various cognitive processing, problem-solving and social situations. Some tests emphasize reasoned plans of action that anticipate consequences of alternative solutions, motor performance in problem-solving situations that require goal-oriented intentions, and regulation of performance for achieving a desired outcome.

Language. Language assessment typically focuses on phonology, morphology, syntax, semantics, and pragmatics. Receptive and expressive language functions may be assessed, including listening, reading, talking, and written language skills and abilities. Assessment of central language disorders focuses on functional speech and verbal comprehension measured through oral, written, or gestural modes; lexical access and elaboration; repetition of spoken language; and associative verbal fluency.

When assessing persons who are non-native English speakers or who are bilingual or

multilingual, language assessment often includes an assessment of language competence and the order of dominance among the different languages. If a multilingual person is assessed for a possible language disorder, one issue for the professional to consider is the degree to which the disorder may be due more directly to language-related qualities (e.g., phonological, morphological, syntactic, semantic, pragmatic delays; mental retardation; peripheral sensory or central neurological impairment; psychological conditions; hearing disorders) than to dominance of a non-English language.

Academic Achievement. Academic achievement tests are measures of academic knowledge and skills that a person has acquired in formal and informal learning opportunities. Two major types of academic achievement tests include general achievement batteries and diagnostic achievement tests. General achievement batteries are designed to assess a person's level of learning in multiple areas (e.g., reading, mathematics, spelling, social studies, science). Diagnostic achievement tests, on the other hand, typically focus on one particular subject area (e.g., reading) and assess important academic skills in greater detail. Test results are used to determine the test taker's strengths as well as specific difficulties and may help identify sources of the difficulties and ways to overcome them. Chapter 13 provides additional detail on academic achievement testing in educational settings.

SOCIAL, ADAPTIVE, AND PROBLEM BEHAVIOR TESTING

Measures of social, adaptive, and problem behaviors assess ability and motivation to care for one's self and to relate to others. Adaptive behaviors include a repertoire of knowledge, skills, and abilities that enable a person to meet the daily demands and expectations of the environment, such as eating, dressing, using transportation, interacting with peers, communicating with others, making purchases, managing money, maintaining a schedule, remaining in school, and maintaining a job.

PART III / PSYCHOLOGICAL TESTING AND ASSESSMENT

Problem behaviors include behavioral adjustment difficulties that interfere with a person's effective functioning in daily life situations.

FAMILY AND COUPLES TESTING

Family testing addresses the issues of family dynamics, cohesion, and interpersonal relations among family members including partners, parents, children, and extended family members. Tests developed to assess families and couples are distinguished by measuring the interaction patterns of partial or whole families, requiring simultaneous focus on two or more family members in terms of their transactions. Testing with couples may address personal factors such as issues of intimacy, compatibility, shared interests, trust, and spiritual beliefs.

PERSONALITY TESTING

Broadly considered, the assessment of personality requires a synthesis of aspects of an individual's functioning that contribute to the formulation and expression of thoughts, attitudes, emotions, and behaviors. In the assessment of an individual, cognitive and emotional functioning may be considered separately, but their influences are interrelated. For example, a person whose perceptions are highly accurate, or who is relatively stable emotionally, may be able to control suspiciousness better than can a person whose perceptions are inaccurate or distorted or who is emotionally unstable.

Scores on a personality test may be regarded as reflecting the underlying theoretical constructs or empirically derived scales or factors that guided the test's construction. The stimulus and response formats of personality tests vary widely. Some include a series of questions (e.g., self-report inventories) to which the test taker is required to choose from several well-defined options; others involve being placed in a novel situation in which the test taker's response is not completely structured (e.g., responding to visual stimuli, telling stories, discussing pictures, or responding to other projective stimuli). The responses are scored and combined into either

logically or statistically derived dimensions established by previous research.

Personality tests may be designed to focus on the assessment of normal or abnormal attitudes, feelings, traits, and related characteristics. Tests intended to measure normal personality characteristics are constructed to yield scores reflecting the degree to which a person manifests personality dimensions empirically identified and hypothesized to be present in the behavior of most individuals. A person's configuration of scores on these dimensions is then used to infer how the person behaves presently and how she/he may behave in new situations. Test scores outside of the expected range may be considered extreme expressions of normal traits or indicative of psychopathology. Such scores also may reflect normal functioning of the person within a culture different from that of the normative population sample.

Other personality tests are designed specifically to measure constructs underlying abnormal functioning and psychopathology. Developers of some of these tests use previously diagnosed individuals to construct their scales and base their inferences on the association between the test's scale scores, within a given range, and the behavioral correlates of persons who scored within that range. If inferences made from scores go beyond the theory that guided the test's construction, then the inferences must be validated by collecting and analyzing additional relevant data.

VOCATIONAL TESTING

Vocational testing generally includes the measurement of interests, work needs, and values, as well as consideration and assessment of related elements of career development, maturity, and indecision. The results from inventories that assess these constructs often are used for enhancing personal growth and understanding, career counseling, outplacement counseling, and vocational decision making. These interventions frequently take place in the context of educational settings.

However, interest inventories and measures of work values also may be used in workplace settings as part of training and development programs, for career planning, or for selection, placement, and advancement decisions.

Interest Inventories. The measurement of interests is designed to identify a person's preferences for various activities. Self-report interest inventories are widely used to assess personal preferences including likes and dislikes for various work and leisure activities, school subjects, occupations, or types of people. The resulting scores may provide insight into types and patterns of differential interests in educational curricula (e.g., college majors), in different fields of work (e.g., specific occupations), or in more general or basic areas of interests related to specific activities (e.g., sales, office practices, or mechanical activities).

Work Values Inventories. The measurement of work values identifies a person's preferences for the various reinforcements one may obtain from work activities. Sometimes these values are identified as needs that persons seek to satisfy. Work values or needs may be categorized as intrinsic and important for the pleasure gained from the activity (e.g., independence, ability utilization, achievement) or as extrinsic and important for the rewards they bring (e.g., coworkers, supervisory relations, working conditions). The format of work values tests usually involves a self-rating of the importance of the value associated with qualities described by the items.

Measures of Career Development, Maturity, and Indecision. Additional areas of vocational assessment include measures of career development and maturity and measures of career indecision. Inventories that measure career development and maturity typically elicit client self-descriptions in response to items that inquire about the individual's knowledge of the world of work; self-appraisal of one's decision-making skills; attitudes toward careers and career choices; and the degree to which the individual already has engaged in career

planning. Measures of career indecision usually are constructed and standardized to assess both the level of career indecision of a client as well as the reasons for, or antecedents of, indecision. Such career development, maturity, and indecision findings may be used with individuals and groups to guide the design and delivery of career services and to evaluate the effectiveness of career interventions.

Purposes of Psychological Testing

For purposes of this chapter, psychological tests have been divided into four categories: testing for diagnosis; intervention planning and outcome evaluation; legal and governmental decisions; and personal awareness, growth and action. However, these categories are not always mutually exclusive.

TESTING FOR DIAGNOSIS

Diagnosis refers to a process that includes the collection and integration of test results with prior and current information about a person together with relevant contextual conditions to identify characteristics of healthy psychological functioning as well as psychological disorders. Disorders may manifest themselves in information obtained during the testing of an individual's cognitive, emotional, social, personality, neuropsychological, physical, perceptual, and motor attributes.

Psychodiagnosis. Psychological tests are helpful to professionals involved in the psychological diagnosis of an individual. Testing may be performed to confirm a hypothesized diagnosis or to rule out alternative diagnoses. Psychodiagnosis is complicated by the prevalence of comorbidity between diagnostic categories. For example, a client diagnosed as suffering from schizophrenia simultaneously may be diagnosed as suffering from depression. Or, a child diagnosed as having a learning disability also may be diagnosed as suffering from an attention deficit disorder. The goal of psychodiagnosis is to assist each client in receiving the appropriate interventions for the psychological or behavioral

PART III / PSYCHOLOGICAL TESTING AND ASSESSMENT

dysfunctions that the client, or a third party, views as impairing the client's expected functioning and/or enjoyment of life. In developing treatment plans, professionals often use non-categorical diagnostic descriptions of client functioning along treatment-relevant dimensions (e.g., degree of anxiety, amount of suspiciousness, openness to interpretations, amount of insight into behaviors, and level of intellectual functioning).

The first step in evaluating a test's suitability to yield scores or information indicative of a particular diagnostic syndrome is to compare the construct that the test is intended to measure with the symptomatology described in the diagnostic criteria. This step is important because different diagnostic systems may use the same diagnostic term to describe different symptoms; even within one diagnostic system the symptoms described by the same term may differ between editions of the manual identifying the diagnostic criteria. Similarly, a test that uses a diagnostic term in its title may differ significantly from another test using a similar title or from a subscale with the same term. For example, some diagnostic systems may define depression by behavioral symptomatology (e.g., psychomotor retardation, disturbance in appetite or sleep) or by affective symptomatology (e.g., dysphoric feeling, emotional flatness) or by cognitive symptomatology (e.g., thoughts of hopelessness, morbidity) or some other symptomatology. Further, rarely are the symptoms of diagnostic categories mutually exclusive. Hence, it can be expected that a given symptom may be shared by several diagnostic categories. More knowledgeable and precisely drawn inferences relating to a diagnosis may be obtained from test scores if appropriate weight is given to the symptoms included in the diagnostic category and to the suitability of each test to assess the symptoms.

Different methods may be used to assess particular diagnostic categories. Some methods rely primarily on structured interviews using a "yes" or "no" format in which the professional

is interested in the presence or absence of diagnosis-specific symptomatology. Other methods often rely principally on tests of personality or cognitive functioning and use configurations of obtained scores. These configurations of scores indicate the degree to which a client's responses are similar to those of individuals who have been determined by prior research to belong to a specific diagnostic group.

Diagnoses made with the help of test scores typically are based on empirically demonstrated relationships between the test score and the diagnostic category. Validity studies that demonstrate relationships between test scores and diagnostic categories currently are available for some diagnostic categories. Sometimes tests that do not have supporting validity studies also may be useful to the professional in arriving at a diagnosis. This also may occur, for example, when the symptoms assessed by a test are a subset of the criteria that comprise a particular diagnostic category. While it often is not feasible for individual professionals to personally conduct research into relationships between obtained scores and inferences, their familiarity with the body of the research literature that examines these relationships is important.

The professional often can enhance the diagnostic inferences derived from test scores by integrating the test results with inferences made from other sources of information regarding the client's functioning such as self-reported history or information provided by significant others or systematic observations in the natural environment or in the testing setting. In arriving at a diagnosis, a professional also looks for information that does not corroborate the diagnosis, and in those instances, places appropriate limits on the degree of confidence placed in the diagnosis. When relevant to the referral issue, the professional acknowledges alternative diagnoses that may require consideration. Particular attention is paid to all relevant available data before concluding that a client falls into a diagnostic category. Cultural sensitivity is paramount to avoid misdiagnosing and over

PSYCHOLOGICAL TESTING AND ASSESSMENT / PART III

pathologizing culturally appropriate behavior, affect or cognition. Tests also are used to assess the appropriateness of continuing the initial diagnostic characterization, especially after a course of treatment or if the client's psychological functioning has changed over time.

Neuropsychodiagnosis. Neuropsychological testing analyzes the current psychological and behavioral status, including manifestations of neurological, neuropathological, and neurochemical changes that may arise during development or from brain injury or illness. The purposes of neuropsychological testing typically include, but are not limited to, the following: differential diagnoses between psychogenic and neurogenic sources of cognitive, perceptual, and personality dysfunction; differential diagnoses between two or more suspected etiologies of cerebral dysfunction; evaluation of impaired functioning secondary to a cerebral, cortical, or subcortical event; establishment of neuropsychological baseline measurements for monitoring progressive cerebral disease or recovery effects; comparison of pre- and post-pharmacologic, surgical, behavioral, or psychological interventions; identification of patterns of higher cortical function and dysfunction for the formulation of rehabilitation strategies and for the design of remedial procedures; and characterizing brain-behavior functions to assist the trier of fact in criminal and civil legal actions.

TESTING FOR INTERVENTION PLANNING AND OUTCOME EVALUATION

Professionals often rely on test results for assistance in planning, executing, and evaluating interventions. Therefore, their awareness of validity information that supports or does not support the relationship between test results, prescribed interventions, and desired outcome is important. Interventions may be intended to prevent the onset of one or more symptoms, to stabilize or overcome them, to ameliorate their effects, to minimize their impact, and to provide for a person's basic physical, psychological, and social needs. Intervention planning typical-

ly occurs following an evaluation of the nature and severity of a disorder and a review of personal and contextual conditions that may impact its resolution. Subsequent evaluations may occur in an effort to diagnose further the nature and severity of the disorder, to review the effects of interventions, to revise them as needed, and to meet ethical and legal standards.

TESTING FOR JUDICIAL AND GOVERNMENTAL DECISIONS

Clients may voluntarily seek psychological testing as part of psychological assessments to assist in matters before a court or other governmental agencies. Conversely, courts or other governmental agencies sometimes require a client to submit involuntarily to a psychological or neuropsychological assessment that may involve a wide range of psychological tests. The goal of these psychological assessments is to provide important information to a third party, client's attorney, opposing attorney, judge, or administrative board about the psychological functioning of the client that has bearing on the legal issues in question. At the outset of evaluations for judicial and government decisions, it is imperative to clarify the purpose of the evaluation, who will have access to the test results and the reports, and any rights that the client may have to refuse to participate in court-ordered evaluations.

The goals of psychological testing in judicial and governmental settings are informed and constrained by the legal issues to be addressed, and a detailed understanding of their salient aspects is essential. Legal issues may arise as part of a civil proceeding (e.g., involuntary commitment, testamentary capacity, competence to stand trial, parole, child custody, personal injury, discrimination issues), a criminal proceeding (e.g., competence to stand trial, not guilty by reason of insanity, mitigating circumstances in sentencing), determination of reasonable accommodations for employees with disabilities, or an administrative proceeding or decision (e.g., license revocation, parole, worker's compensation). Each of these legal issues is

PART III / PSYCHOLOGICAL TESTING AND ASSESSMENT

defined in law applicable to a particular legislative jurisdiction. The definition of each legal issue may be jurisdiction specific. For example, the criteria by which a person can be involuntarily committed often differ between legislative jurisdictions. Furthermore, tests initially administered for one purpose also may be used for another purpose (e.g., initially used for a civil case but later used in administrative or criminal proceedings).

Legislatures, courts, and other administrative bodies often define legal issues in commonly used language, not in diagnostic or other technical psychological terms. The professional is responsible for explaining the diagnostic frame of reference, including test scores and inferences made from them, in terms of the legal criteria by which the jury, judge, or administrative board will decide the legal issue. For example, a diagnosis of schizophrenia or neuropsychological impairment, which does not also include a reference to the legal criteria, neither precludes an examinee from obtaining sole custody of children in a child custody dispute nor does it necessarily acquit a person of criminal responsibility.

In instances involving legal or quasi-legal issues, it is important to assess the examinee's test-taking orientation including response bias to ensure that the legal proceedings have not affected the responses given. For example, a person seeking to obtain the greatest possible monetary award for a personal injury may be motivated to exaggerate cognitive and emotional symptoms, while persons attempting to forestall the loss of a professional license may attempt to portray themselves in the best possible light by minimizing symptoms or deficits. In forming an assessment opinion, it is necessary to interpret the test scores with informed knowledge relating to the available validity and reliability evidence. When forming such opinions, it also is necessary to integrate a client's test scores with all other sources of information that bear on current status including psychological, medical, educational, occupational, legal, and other relevant collateral records.

Some tests are intended to provide information about a client's functioning that helps clarify a given legal issue (e.g., parental functioning in a child custody case or ability to understand charges against a defendant in competency to stand trial matters). The manuals of some tests also provide demographic and actuarial data for normative groups that are representative of persons involved in the legal system. However, many tests measure constructs that are generally relevant to the legal issues even though norms specific to the judicial or governmental context may not be available. Professionals are expected to make every effort to be aware of evidence of validity and reliability that supports or does not support their inferences and to place appropriate limits on the opinions rendered. Test users who practice in judicial and government settings are expected to be aware of conflicts of interest that may lead to bias in the interpretation of test results.

Protecting the confidentiality of a client's test results and of the test instrument itself poses particular challenges for professionals involved with attorneys, judges, jurors, and other legal and quasi-legal decision makers. The test taker does have a right to expect that test results will be communicated only to persons who are legally authorized to receive them and that other information from the testing session that is not relevant to the evaluation will not be reported. It is important for the professional to be apprised of possible threats to confidentiality and test security (e.g., releasing the test questions, the examinee's responses, and raw and scaled scores on tests to another qualified professional) and to seek, if necessary, appropriate legal and professional remedies.

TESTING FOR PERSONAL AWARENESS, GROWTH, AND ACTION

Tests and inventories frequently are used to provide information to help individuals to understand themselves, to identify their own strengths and weaknesses, and to otherwise clarify issues important to their own decision

making and development. For example, test results from personality inventories may help clients better understand themselves and also understand their interactions with others. Results from interest inventories and tests of ability may be useful to individuals who are making educational and career decisions. Appropriate cognitive and neuropsychological tests that have been normed and standardized for children may facilitate the monitoring of development and growth during the formative years when relevant interventions may be more efficacious for preventing potentially disabling learning disabilities from being overlooked or misdiagnosed.

Test results may be used for self-exploration, self-growth, and decision making in several ways. First, the results can provide individuals with new information that allows them to compare themselves with others or to evaluate themselves by focusing on self-descriptions and characterizations. Test results also may serve to stimulate discussions between a client and professional, to facilitate client insights, to provide directions for future considerations, to help individuals identify strengths and assets, and to provide the professional with a general framework for organizing and integrating information about an individual. Testing for personal growth may take place in training and development programs, within an educational curriculum, during psychotherapy, in rehabilitation programs as part of an educational or career planning process, or in other situations.

Summary

The application of psychological tests continues to expand in scope and depth on a course that is characterized by an increasingly diverse set of purposes, procedures, and assessment needs and challenges. Therefore, the responsible use of tests in practice requires a commitment by the professional to develop and maintain the necessary knowledge and competence to select, administer, and interpret tests and inventories

as crucial elements of the psychological testing and assessment process. The standards in this chapter provide a framework for guiding the professional toward achieving relevance and effectiveness in the use of psychological tests within the boundaries or limits defined by the professional's educational, experiential and ethical foundations. Earlier chapters and standards that are relevant to psychological testing and assessment describe general aspects of test quality (chapters 1-6, chapter 11), test fairness (chapters 7-10), and test use (chapter 11). Chapter 13 discusses educational applications; chapter 14 discusses test use in the workplace, including credentialing, and the importance of collecting data that provide evidence of a test's accuracy for predicting job performance; and chapter 15 discusses test use in program evaluation and public policy.

Standard 12.1

Those who use psychological tests should confine their testing and related assessment activities to their areas of competence, as demonstrated through education, supervised training, experience, and appropriate credentialing.

Comment: The responsible use and interpretation of test scores require appropriate levels of experience and sound professional judgment. Competency also requires sufficient familiarity with the population from which the test taker comes to allow appropriate interaction, test selection, test administration, and test interpretation. For example, when personality tests and neuropsychological tests are administered as part of a psychological assessment of an individual, the test scores must be understood in the context of the individual's physical and emotional state, as well as the individual's cultural, educational, occupational, and medical background, and must take into account other evidence relevant to the tests used. Test interpretation in this context requires professionally responsible judgment that is exercised within the boundaries of knowledge and skill afforded by the professional's education, training, and supervised experience.

Standard 12.2

Those who select tests and interpret test results should refrain from introducing biases that accommodate individuals or groups with a vested interest in decisions affected by the test interpretation.

Comment: Individuals or groups with a vested interest in the significance or meaning of the findings from psychological testing include many school personnel, attorneys, referring health professionals, employers, professional associates, and managed care organizations. In some settings a professional may have a professional relationship with multiple clients (e.g.,

with both the test taker and the organization requesting assessment). A professional engaged in a professional relationship with multiple clients takes care to ensure that the multiple relationships do not become a conflict of interest that would occur when the professional's judgment toward one client is unduly influenced by his or her relationship with the other client. Test selections and interpretations that favor a special external expectation or perspective by deviating from established principles of sound test interpretation are unprofessional and unethical.

Standard 12.3

Tests selected for use in individual testing should be suitable for the characteristics and background of the test taker.

Comment: Considerations for test selection should include culture, language and/or physical requirements of the test and the availability of norms and evidence of validity for a population representative of the test taker. If no normative or validity studies are available for the population at issue, test interpretations should be qualified and presented as hypotheses rather than conclusions.

Standard 12.4

If a publisher suggests that tests are to be used in combination with one another, the professional should review the evidence on which the procedures for combining tests is based and determine the rationale for the specific combination of tests and the justification of the interpretation based on the combined scores.

Comment: For example, if measures of developed abilities (e.g., achievement or specific or general abilities) or personality are packaged with interest measures to suggest a requisite combination of scores, or a neuropsychological battery is being applied, then supporting validity data for such combinations of scores should be available.

STANDARDS

PSYCHOLOGICAL TESTING AND ASSESSMENT / PART III

Standard 12.5

The selection of a combination of tests to address a complex diagnosis should be appropriate for the purposes of the assessment as determined by available evidence of validity. The professional's educational training and supervised experience also should be commensurate with the test user qualifications required to administer and interpret the selected tests.

Comment: For example, in a neuropsychological assessment for evidence of an injury to a particular area of the brain, it is necessary to select a combination of tests of known diagnostic sensitivity and specificity to impairments arising from trauma to various regions of the cerebral hemispheres.

Standard 12.6

When differential diagnosis is needed, the professional should choose, if possible, a test for which there is evidence of the test's ability to distinguish between the two or more diagnostic groups of concern rather than merely to distinguish abnormal cases from the general population.

Comment: Professionals will find it particularly helpful if evidence of validity is in a form that enables them to determine how much confidence can be placed in inferences regarding an individual. Differences between group means and their statistical significance provide inadequate information regarding validity for individual diagnostic purposes. Additional information might consist of confidence intervals, effect sizes, or a table showing the degree of overlap of predictor distributions among different criterion groups.

Standard 12.7

When the validity of a diagnosis is appraised by evaluating the level of agreement between test-based inferences and the diagnosis, the

diagnostic terms or categories employed should be carefully defined or identified.

Standard 12.8

Professionals should ensure that persons under their supervision, who administer and score tests, are adequately trained in the settings in which the testing occurs and with the populations served.

Standard 12.9

Professionals responsible for supervising group testing programs should ensure that the individuals who interpret the test scores are properly instructed in the appropriate methods for interpreting them.

Comment: If, for example, interest inventories are given to college students for use in academic advising, the professional who supervises the academic advisors is responsible for ensuring that the advisors know how to provide an examinee an appropriate interpretation of the test results.

Standard 12.10

Prior to testing, professionals and test administrators should provide the test taker with appropriate introductory information in language understandable to the test taker. The test taker who inquires also should be advised of opportunities and circumstances, if any, for retesting.

Comment: The client should understand testing time limits, who will have access to the test results, if and when test results will be shared with the test taker, and if and when decisions based on the test results will be shared with the test taker.

Standard 12.11

Professionals and others who have access to test materials and test results should ensure

the confidentiality of the test results and testing materials consistent with legal and professional ethics requirements.

Comment: Professionals should be knowledgeable and conform to record-keeping and confidentiality guidelines required by the state or province in which they practice and the professional organizations to which they belong. Confidentiality has different meanings for the test developer, the test user, the test taker, and third parties (e.g., school, court, employer). To the extent possible, the professional who uses tests is responsible for managing the confidentiality of test information across all parties. It is important for the professional to be aware of possible threats to confidentiality and the legal and professional remedies available. Professionals also are responsible for maintaining the security of testing materials and for protecting the copyrights of all tests to the extent permitted by law.

Standard 12.12

The professional examines available norms and follows administration instructions, including calibration of technical equipment, verification of scoring accuracy and replicability, and provision of settings for testing that facilitate optimal performance of test takers. However, in those instances where realistic rather than optimal test settings will best satisfy the assessment purpose, the professional should report the reason for using such a setting and, when possible, also conduct the testing under optimal conditions to provide a comparison.

Comment: Because the normative data against which a client's performance will be evaluated were collected under the reported standard procedures, the professional needs to be aware of and take into account the effect that non-standard procedures may have on the client's obtained score. When the professional uses

tests that employ an unstructured response format, such as some projective techniques and informal behavioral ratings, the professional should follow objective scoring criteria, where available and appropriate, that are clear and minimize the need for the scorer to rely only on individual judgment. The testing may be conducted in a realistic, less than optimal, setting to determine how a client with an attentional disorder, for example, performs in a noisy or distracting environment rather than in an optimal environment that typically protects the test taker from such external threats to performance efficiency.

Standard 12.13

Those who select tests and draw inferences from test scores should be familiar with the relevant evidence of validity and reliability for tests and inventories used and should be prepared to articulate a logical analysis that supports all facets of the assessment and the inferences made from the assessment.

Comment: A presentation and analysis of validity and reliability evidence generally is not needed in a written report, because it is too cumbersome and of little interest to most report readers. However, in situations in which the selection of tests may be problematic (e.g., verbal subtests with deaf clients), a brief description of the rationale for using or not using particular measures is advisable.

When potential inferences derived from psychological test data are not supported by evidence of validity yet may hold promise for future validation, they may be described by the test developer and professional as hypotheses for further validation in test interpretation. Such interpretive remarks should be qualified to communicate to the source of the referral that such inferences do not as yet have adequately demonstrated evidence of validity and should not be the basis for a diagnostic decision or prognostic formulation.

STANDARDS

PSYCHOLOGICAL TESTING AND ASSESSMENT / PART III

Standard 12.14

The interpretation of test results in the assessment process should be informed when possible by an analysis of stylistic and other qualitative features of test-taking behavior that are inferred from observations during interviews and testing and from historical information.

Comment: Such features of test-taking behavior include manifestations of fatigue, momentary fluctuations in emotional state, rapport with the examiner, test taker's level of motivation, withholding or distortion of response as seen in instances of deception and malingering or in instances of pseudoneurological conditions, and unusual response or general adaptation to the testing environment.

Standard 12.15

Those who use computer-generated interpretations of test data should evaluate the quality of the interpretations and, when possible, the relevance and appropriateness of the norms upon which the interpretations are based.

Comment: Efforts to reduce a complex set of data into computer-generated interpretations of a given construct may yield grossly misleading or simplified analyses of meanings of test scores, that in turn may lead to faulty diagnostic and prognostic decisions as well as mislead the trier of fact in judicial and government settings.

Standard 12.16

Test interpretations should not imply that empirical evidence exists for a relationship among particular test results, prescribed interventions, and desired outcomes, unless empirical evidence is available for populations similar to those representative of the examinee.

Standard 12.17

Criterion-related evidence of validity should be available when recommendations or decisions are presented by the professional as having an actuarial basis.

Standard 12.18

The interpretation of test or test battery results generally should be based upon multiple sources of convergent test and collateral data and an understanding of the normative, empirical, and theoretical foundations as well as the limitations of such tests.

Comment: A given pattern of test performances represents a cross-sectional view of the individual being assessed within a particular context (i.e., medical, psychosocial, educational, vocational, cultural, ethnic, gender, familial, genetic, and behavioral). The interpretation of findings derived from a complex battery of tests in such contexts requires appropriate education, supervised experience, and an appreciation of procedural, theoretical, and empirical limitations of the tests.

Standard 12.19

The interpretation of test scores or patterns of test battery results should take cognizance of the many factors that may influence a particular testing outcome. Where appropriate, a description and analysis of the alternative hypotheses or explanations that may have contributed to the pattern of results should be included in the report.

Comment: Many factors (e.g., unusual testing conditions, motivation, educational level, employment status, lateral sensorimotor usage preferences, health, or disability status) may influence individual testing results. When such factors are known to introduce construct-irrelevant variance in component test scores, those factors should be considered during test score interpretations.

Standard 12.20

Except for some judicial or governmental referrals, or in some employment testing situations when the client is the employer, professionals should share test results and interpretations with the test taker. Such information should be expressed in language that the test taker, or when appropriate the test taker's legal representative, can understand.

Comment: For example, in rehabilitation settings, where clients typically are required to participate actively in intervention programs, sharing of such information, expressed in terms that can be understood readily by the client and family members, may facilitate the effectiveness of intervention.

13. EDUCATIONAL TESTING AND ASSESSMENT

Background

This chapter concerns testing in formal educational settings from kindergarten through post-graduate training. Results of tests administered to students are used to make judgments, for example, about the status, progress, or accomplishments of individuals or groups. Tests that provide information about individual performance are used to (a) evaluate a student's overall achievement and growth in a content domain, (b) diagnose student strengths and weaknesses in and across content domains, (c) plan educational interventions and to design individualized instructional plans, (d) place students in appropriate educational programs, (e) select applicants into programs with limited enrollment, and (f) certify individual achievement or qualifications. Tests that provide information about the status, progress, or accomplishments of groups such as schools, school districts, or states are used (a) to judge and monitor the quality of educational programs for all or for particular subsets of individuals, and (b) to infer the success of policies and interventions that have been selected for evaluation. These testing purposes are typically mandated by institutions such as schools and colleges and by governing bodies of public and privately administered educational programs.

In this chapter, three broad areas of educational testing are considered that encompass one or more of the above purposes: (a) routine school, district, state, or other system-wide testing programs; (b) testing for selection in higher education; and (c) individualized and special needs testing. While the second and third areas refer to relatively specific purposes of testing, system-wide testing programs can encompass multiple individual and group purposes. For each of these areas, the chapter elaborates on the specific purposes and domains encompassed and raises specific issues of tech-

nical quality and fairness in testing that may not be addressed or emphasized in the preceding chapters. This chapter does not explicitly address issues related to tests constructed and administered by teachers for their own classroom use or provided by publishers of instructional materials. While many aspects of the *Standards*, particularly those in the areas of validity, reliability, test development, and fairness, are relevant to such tests, this document is not intended for tests used by teachers for their own classroom purposes.

Issues in Educational Testing

This chapter first considers some cross-cutting issues: the distinctions among types of tests, the design or use of tests to serve multiple purposes including the measurement of change, and the "stakes" associated with different purposes for testing in education.

DISTINCTIONS AMONG TYPES OF TESTS AND ASSESSMENTS

Tests used in educational settings range from tests consisting of traditional item formats such as multiple-choice items to performance assessments including scorable portfolios. Every test, regardless of its format, measures test-taker performance in a specified domain. Performance assessments, however, attempt to emulate the context or conditions in which the intended knowledge or skills are actually applied. As discussed in chapter 3, they are diverse in nature and can be product-based as well as behavior-based. The execution of the tasks posed in these tests often involves relatively extended time periods, ranging from a few minutes to a class period or more to several hours or days. Examples of such performances might include solving problems using manipulable materials, making complex inferences after collecting information, or explaining orally or in writing

the rationale for a particular course of government action under given economic conditions. The performance task may be undertaken by a single individual or a team of students. Performance assessments may require increased testing time to provide sufficient domain sampling for reasonable estimates of individual attainment and for making generalizations to the broader domain. Extended time periods, collaboration, and the use of ancillary materials pose great challenges to the standardization of administration and scoring of some performance assessments. This is particularly true when test takers define their own tasks or when they select their own work products for evaluation. When this is the case, test takers need to be aware of the basis for scoring as well as the nature of the criteria that will be applied. Further, performance assessments often require complex procedures and training to increase the accuracy of judgments made by those evaluating student performance (see chapter 3).

An individual portfolio may be used as another type of performance assessment. Scorable portfolios are systematic collections of educational products typically collected over time and possibly amended over time. The particular purpose of the portfolio determines whether it will include representative products, the best work of the student, or indicators of progress. The purpose also dictates who will be responsible for compiling the contents of the portfolio—the examiner, the student, or both parties working together. The more standardized the contents and procedures of administration, the easier it is to establish comparability of portfolio-based scores. Establishing comparability requires portfolios to be constructed according to test specifications and standards, and the development of objective procedures to judge their quality. The test specifications for portfolios may indicate that students are to make certain decisions about the nature of the work to be included. For example, in constructing an art portfolio, students may select the media that best represent their work. Establishing comparability

also requires specifications regarding the kinds of assistance the student may have received during portfolio preparation. It is particularly difficult to compare the performance of students whose portfolios may vary in content. All performance assessments, including scorable portfolios, are judged by the same standards of technical quality as traditional tests of achievement.

Electronic media are often used both to present testing material and to record and score test takers' responses. These tests may be administered in schools, in special laboratory settings, or in external testing centers. Examples include simple enhancements of text by audio-taped instructions to facilitate student understanding, computer-based tests traditionally given in paper-and-pencil format, computer-adaptive tests, and newer, interactive multimedia testing situations where attributes of performance assessments are supported by computer. Some computer-based tests also may have the capacity to capture aspects of students' processes as they solve test items. They may, for example, monitor time spent on items, solutions tried and rejected, or editing sequences for texts. Electronic media also make it possible to provide test administration conditions designed to assist students with particular needs, such as those with different language backgrounds, attention problems, or physical disabilities. Computers can also help identify the contributions of individuals to a group task completed by a team or in geographically remote locations on a network.

Computer-based tests are evaluated by the same technical quality standards as other tests administered through more traditional means. It is especially important that test takers be familiarized with the media of the test so that any unfamiliarity with computers or strategies does not lead to inferences based on construct-irrelevant variance. Furthermore, it is important to describe scoring algorithms, expert models upon which they may be based, and technical data supporting their use in any documentation accompanying the testing system. It is important, however, to assure that the docu-

PART III / EDUCATIONAL TESTING AND ASSESSMENT

mentation does not jeopardize the security of the items that could adversely affect the validity of score interpretations. Some computer-based tests may also generate recommendations for instructional practices based on test results. Describing the basis for these recommendations assists the user in evaluating their applicability in a given situation.

MULTIPLE PURPOSES AND MEASURING CHANGE

Many tests are designed or used to serve multiple purposes in education. For example, a test may be used to monitor individual student achievement as well as to evaluate the quality of educational programs at the school or district level. As another example, a test may be used to evaluate an individual's performance relative to the performance of one or more reference populations as well as to evaluate the level of the individual's competence in some defined domain (see chapters 3 and 4). The evidence needed for the technical quality of one purpose, however, will differ from the evidence needed for another purpose. Consequently, it is important to evaluate the evidence of technical quality for each purpose of testing.

Test results may be used to infer the growth or progress as well as the status of individuals or groups of students, such as when tests are expected to reveal the effects of instruction, of changes in educational policy, or of other interventions. In such cases, the test's ability to detect change is essential. If differences in scores are reported, the technical quality of the differences needs attention. More generally, whenever inferences about growth or progress are made, it is important to evaluate the validity of those inferences.

STAKES OF TESTING

The importance of the results of testing programs for individuals, institutions, or groups is often referred to as the *stakes* of the testing program. At the individual level, when significant educational paths or choices of an individual are directly affected by test performance, such as

whether a student is promoted or retained at a grade level, graduated, or admitted or placed into a desired program, the test use is said to have high stakes. A low-stakes test, on the other hand, is one administered for informational purposes or for highly tentative judgments such as when test results provide feedback to students, teachers, and parents on student progress during an academic period. Testing programs for institutions can have high stakes when aggregate performance of a sample or of the entire population of test takers is used to infer the quality of service provided, and decisions are made about institutional status, rewards, or sanctions based on test results. For example, the quality of reading curriculum and instruction may be judged on the basis of test results because test scores can indicate the rate of student progress or the levels of attainment reached by groups of students. Even when test results are reported in the aggregate and intended for a low-stakes purpose such as monitoring the educational system, the public release of data can raise the stakes for particular schools or districts. Judgments about program quality, personnel, and educational programs might be made and policy decisions might be affected, even though the tests were not intended or designed for those purposes.

The higher the stakes associated with a given test use, the more important it is that test-based inferences are supported with strong evidence of technical quality. In particular, when the stakes for an individual are high, and important decisions depend substantially on test performance, the test needs to exhibit higher standards of technical quality for its avowed purposes than might be expected of tests used for lower-stakes purposes (see chapters 1, 2, and 7 for a more thorough discussion on validity, reliability, and bias in testing, respectively). Although it is never possible to achieve perfect accuracy in describing an individual's performance, efforts need to be made to minimize errors in estimating individual scores or in classifying individuals in pass/fail or admit/reject categories.

Further, enhancing validity for high-stakes purposes, whether individual or institutional, typically entails collecting sound collateral information both to assist in understanding the factors that contributed to test results and to provide corroborating evidence that supports inferences based on test results. These issues will be addressed more fully as they relate to the three areas of testing described below.

School, District, State, or Other System-Wide Testing Programs

As indicated previously, system-wide testing programs can span multiple purposes. At the individual level, tests are used for low-stakes purposes, such as monitoring and providing feedback on student progress, and for more high-stakes purposes, such as certifying students' acquisition of particular knowledge and skills for promotion, placement into special instructional programs, or graduation. At the school, district, state, or other aggregate level, a common purpose of tests is to evaluate the progress made by groups of students or to monitor the long-term effectiveness of the overall educational system. Educational testing programs may also permit comparisons among the performance of various groups of students in different programs or in diverse settings for the purpose of making an evaluation of those learning environments. Chapter 15 provides a more thorough discussion on program evaluation.

In these contexts, educational tests are designed to measure certain aspects of students' knowledge and skills as reflected in curriculum goals and standards. There may be considerable variation in the breadth and depth of the knowledge and skills that are measured by such tests. Some educational tests focus on the test takers' general ability or knowledge in a particular content area, such as their understanding of mathematics or science. Other tests focus on test takers' specific knowledge of a topic in detail, such as trigonometry.

Still others emphasize specific skills or procedures, such as the ability to write persuasively or to design, conduct, and interpret the results of a scientific experiment. Tests may address other cognitive aspects of test takers' development, such as their ability to work with others to solve problems or their self-reported habits and attitudes, as well as noncognitive aspects, such as students' ability to perform particular physical tasks. In most cases, valid interpretation of the results requires that evidence of the fit between the test domain and the relevant curriculum goals or standards be ascertained.

Testing programs may involve the use of tests designed to represent a set of general educational standards as determined for instance by the state, district, or relevant educational professional organization. Such tests are conceptually similar to criterion-referenced tests, in that a set of content standards is developed that is intended to provide broad specifications for student performance by delimiting the content and general skills to be measured. Subsequently, descriptive or empirical targets or levels of achievement are developed and referred to as performance standards. These performance standards are intended to define further the knowledge and skills required of students for each of the different categories of proficiency.

This type of testing may involve the development of a new test to assess the relevant content and skills or the selection of an existing test that can be referenced to the standards. Whether a test is designed or selected, valid interpretation of the results in light of the standards entails assessment of the degree of fit between the test domain and contents and the descriptive statements of standards or goals. This involves a process of mapping or referencing the content and skills of the test to those of the standards to be sure that gaps or imbalances do not occur. The curriculum goals or standards may be sufficiently broad to encompass many different ways for students to demonstrate their status, accomplishments, or

PART III / EDUCATIONAL TESTING AND ASSESSMENT

progress. Moreover, some goals or standards may not lend themselves to conventional test formats. These are cases in which the test may result in construct underrepresentation that refers to the extent to which a test fails to capture important aspects of what it is intended to measure. Chapter 1 provides a more thorough discussion of construct underrepresentation. In these cases, interpretation of test results in light of goals or standards is enhanced by an understanding of what is not covered as well as what is covered by the test. Sometimes, additional commercial or locally developed tests are administered within a particular jurisdiction, and attempts are made to link these existing tests to the proficiency levels reported for the new test or to provide other evidence of comparability. It is important to provide logical and empirical validity evidence of any reported links. For example, evidence can be collected to determine the extent to which the existing test can provide information about the proficiency of individual students and groups of students in the particular content areas and skills addressed by the standards. The validity of such links is problematic to the extent that the tests measure different content (see chapter 4 for a discussion on issues in equating and linking tests).

When inferences are to be drawn about the performance of groups of students, practical considerations and the format of the test (e.g., performance assessment) often dictate that different subgroups of students within each unit respond to different sets of tasks or items, a procedure referred to as matrix sampling. This matrix sampling approach allows for a test to better represent the breadth of the target domain without increasing the testing time for each test taker. Group-level results are most useful when testing programs and student populations remain sufficiently stable to provide information about trends over time. When a testing program is designed for group-level reporting and employs matrix sampling, reporting individual scores generally is not appropriate.

When interpreting and using scores about individuals or groups of students, consideration of relevant collateral information can enhance the validity of the interpretation, by providing corroborating evidence or evidence that helps explain student performance. Test results can be influenced by multiple factors, including insitutional and individual factors such as the quality of education provided, students' exposure to education (e.g., through regular school attendance), and students' motivation to perform well on the test.

As the stakes of testing increase for individual students, the importance of considering additional evidence to document the validity of score interpretations and the fairness in testing increases accordingly. The validity of individual interpretations can be enhanced by taking into account other relevant information about individual students before making important decisions. It is important to consider the soundness and relevance of any collateral information or evidence used in conjunction with test scores for making educational decisions. Further, fairness in testing can be enhanced through careful consideration of conditions that affect students' opportunities to demonstrate their capabilities. For example, when tests are used for promotion and graduation, the fairness of individual interpretations can be enhanced by (a) providing students with multiple opportunities to demonstrate their capabilities through repeated testing with alternate forms or through other construct-equivalent means, (b) ensuring students have had adequate notice of skills and content to be tested along with other appropriate test preparation material, (c) providing students with curriculum and instruction that affords them the opportunity to learn the content and skills that are tested, and (d) providing students with equal access to any specific preparation for test taking (e.g., test-taking strategies). Chapter 7 provides a more thorough discussion on fairness in testing.

Collateral information can also enhance interpretation and decisions at the institutional

level. For instance, changes in test scores from year to year may not only reflect changes in the capabilities of students but also changes in the student population (e.g., successive cohorts of students). Differences in scores across ethnic groups may be confounded with differences in socioeconomic status of the communities in which they live and, hence, the educational resources to which students have access. Differences in scores from school to school may similarly reflect differences in resources and activities such as the qualification of teachers or the number of advanced course offerings. While local empirical evidence of the influence of these factors may not be readily available, consideration of evidence from similar contexts available in published literature can enhance the quality of the interpretation and use of current results.

Because public participation is an integral part of educational governance, policymakers, professional educators, and members of the public are concerned with the nature of educational tests, the domains that the tests are intended to measure, the choices in test design, adoption, and implementation, and the issues associated with valid interpretation and uses of test results. It is important that test results be reported in a way that all stakeholders can understand, that enables sound interpretations, and that decreases the chance of misinterpretations and inappropriate decisions.

Large-scale testing is increasingly viewed as a tool of educational policy. From this perspective, tests used for program evaluation, such as some state tests that are aligned to the state's own curriculum standards, are not used solely as measures of school outcomes (see chapter 15 for a more thorough discussion on the use of tests for program evaluation). They are also viewed as a means to influence curriculum and instruction, to hold teachers and school administrators accountable, to increase student motivation, and to communicate performance expectations to students, to teachers, and to the public. If such goals are set forth as

part of the rationale for a testing program, the validity of the testing program needs to be examined with respect to these goals. Beyond any intended policy goals, it is important to consider potential unintended effects that may result from large-scale testing programs. Concerns have been raised, for instance, about narrowing the curriculum to focus only on the objectives tested, restricting the range of instructional approaches to correspond to the testing format, increasing the number of dropouts among students who do not pass the test, and encouraging other instructional or administrative practices that may raise test scores without affecting the quality of education. It is important for those who mandate tests to consider and monitor their consequences and to identify and minimize the potential of negative consequences.

Selection in Higher Education

It is widely recognized that tests are used in the selection of applicants for admission to particular educational programs, especially admissions to colleges, universities, and professional schools. Selection criteria may vary within an institution by academic specialization. In addition to scores from selection tests, many other sources of evidence are used in making selection decisions, including past academic records, transcripts, and grade-point average or rank in class. Scores on tests used to certify students for high school graduation may be used in the college admissions process. Other measures used by some institutions are samples of previous work by students, lists of academic and service accomplishments, letters of recommendation, and student-composed statements evaluated for the appropriateness of the goals and experience of the student or for writing proficiency.

Two major points may be made about the role of tests in the admissions process. Often, scores are used in combination with other sources of information. Some of these supple-

PART III / EDUCATIONAL TESTING AND ASSESSMENT

mental sources of evidence may not be reliably assessed or may lack comparability from applicant to applicant. For this reason, it is important that studies be conducted examining the relationships among test scores, data from other sources of information, and college performance. Second, the public and policymakers are to be cautious about the widespread use of reports of college admission test scores to infer the effectiveness of middle school and high school as well as to compare schools or states. Admissions tests, whether they are intended to measure achievement or ability, are not directly linked to a particular instructional curriculum and, therefore, are not appropriate for detecting changes in middle school or high school performance. Because of differential motivational factors and other demographic variables found across and within pre-collegiate programs, self-selection precludes general comparisons of test scores across demographic groups. Therefore, self-selection also precludes comparisons of test scores among the full ranges of pre-collegiate programs.

Individualized and Special Needs Testing

Individually administered tests are used by school psychologists and other professionals in schools and other related settings to facilitate the learning and development of students who may have special educational needs (see chapter 12). Some of these services are reserved for those students who have gifted capabilities as well as for those students who may have relatively minor academic difficulties (e.g., such as those requiring remedial reading). Other services are reserved for students who display behavioral, emotional, physical, and/or more severe learning difficulties. Services may be provided to students who are in regular classroom settings as well as to students who need more specialized instruction outside of the regular classroom. The ultimate purpose of these services is to

assure all students are placed into appropriate educational programs.

Individually administered tests can serve a number of purposes, including screening, diagnostic classification, intervention planning, and program evaluation. For screening purposes, tests are administered to identify students who might differ significantly from their peers and might require additional assessment. For example, screening tests may be used to identify young children who show signs of developmental disorders and to signal the need for further evaluation. For diagnostic purposes, tests may be used to clarify the types and extent of an individual's difficulties or problems in light of well-established criteria. Test results provide an important basis for determining whether the student meets eligibility requirements for special education and other related services and, if so, the specific types of services that the student needs. Test results may be used for intervention purposes in establishing behavior and learning goals and objectives for the student, planning instructional strategies that should be used, and specifying the appropriate setting in which the special services are to be delivered (e.g., regular classroom, resource room, full-time special class, etc.). Subsequent to the student's placement in special services, tests may be administered to monitor the progress of the student toward prescribed learning goals and objectives. Test results may be used also to evaluate the effectiveness of instruction to determine whether the special services need to be continued, modified, or discontinued.

Many types of tests are used in individualized and special needs testing. These include tests of cognitive abilities, academic achievement, learning processes, visual and auditory memory, speech and language, vision and hearing, and behavior and personality. These tests are used typically in conjunction with other assessment methods such as interviews, behavioral observation, and review of records. Each of these may provide useful data for mak-

ing appropriate decisions about a student. In addition, procedures that aim to link assessment closely to intervention may be used, including behavioral assessments, assessments of learning environments, curriculum-based tests, and portfolios. Regardless of the qualities being assessed and types of data collection methods employed, assessment data used in making special education decisions are evaluated in terms of validity, reliability, and relevance to the specific needs of the students. They must also be judged in terms of their usefulness for designing appropriate educational programs for students who have special needs.

The amount and complexity of the assessment data required for making various decisions about a student will vary depending on the purpose of testing, the needs of the student, and other information already available about the student (e.g., current scores on a relevant test may be on file for some students but not for others). In general, testing for screening and program evaluation purposes typically involves the use of one or two tests rather than comprehensive test batteries. For determining eligibility and designing intervention, testing and assessment is more comprehensive and may involve multiple procedures and sources. Moreover, in-depth analyses and interpretation of the data are necessary.

In special education, tests are selected, administered, and interpreted by school psychologists, school counselors, regular and special educators, speech pathologists, and physical therapists, among other professionals. The validity of inferences will be enhanced if test users possess adequate knowledge of the principles of measurement and evaluation. However, this diverse group of test users may differ in their levels of technical expertise in measurement and degree of professional training in assessment procedures. It is important that professional evaluators administer and interpret only those tests with which they

have training and competence, in order to prevent misuse of tests.

State and federal law generally requires that students who are referred for possible special education services be screened for eligibility. The screening or initial assessment may in turn call for a more comprehensive evaluation. But the large numbers of students to be tested, the high cost of special education programs, and the limits of time create pressures on special education assessment practices. Assessment usually must be completed within a specific number of working days after referral, and, in most instances, the school district is responsible for funding special services recommended by the child study team. Occasionally, administrators might be inclined to use less expensive, less time-consuming, or more readily available testing procedures than a professional evaluator believes are warranted. An example would be the inappropriate use of available, but less adequately trained, staff to evaluate students. There also might be pressures to minimize or overlook problems that require expensive services. These conditions are likely to adversely affect the validity of the interpretation of test results. Adherence to professional standards governing test use in conducting special education assessments is important, in the face of pressures to use more expedient procedures. The responsible use of tests by school personnel can improve the opportunities for promoting the development and learning of all children.

Standard 13.1

When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test user.

Comment: Mandated testing programs are often justified in terms of their potential benefits for teaching and learning. Concerns have been raised about the potential negative impact of mandated testing programs, particularly when they result directly in important decisions for individuals or institutions. Frequent concerns include narrowing the curriculum to focus only on the objectives tested, increasing the number of dropouts among students who do not pass the test, or encouraging other instructional or administrative practices simply designed to raise test scores rather than to affect the quality of education.

Standard 13.2

In educational settings, when a test is designed or used to serve multiple purposes, evidence of the test's technical quality should be provided for each purpose.

Comment: In educational testing, it has become common practice to use the same test for multiple purposes (e.g., monitoring achievement of individual students, providing information to assist in instructional planning for individuals or groups of students, evaluating schools or districts). No test will serve all purposes equally well. Choices in test development and evaluation that enhance validity for one purpose may

diminish validity for other purposes. Different purposes require somewhat different kinds of technical evidence, and appropriate evidence of technical quality for each purpose should be provided by the test developer. If the test user wishes to use the test for a purpose not supported by the available evidence, it is incumbent on the user to provide the necessary additional evidence (see chapter 1).

Standard 13.3

When a test is used as an indicator of achievement in an instructional domain or with respect to specified curriculum standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both tested and target domains should be described in sufficient detail so their relationship can be evaluated. The analyses should make explicit those aspects of the target domain that the test represents as well as those aspects that it fails to represent.

Comment: Increasingly, tests are being developed to monitor progress of individuals and groups toward local, state, or professional curriculum standards. Rarely can a single test cover the full range of performances reflected in the curriculum standards. To assure appropriate interpretations of test scores as indicators of performance on these standards, it is essential to document and evaluate both the relevance of the test to the standards and the extent to which the test represents the standards. When existing tests are selected by a school, district, or state to represent local curricula, it is incumbent on the user to provide the necessary evidence of the congruency of the curriculum domain and the test content. Further, conducting studies of the cognitive strategies and skills employed by test takers or studies of the

STANDARDS

EDUCATIONAL TESTING AND ASSESSMENT / PART III

relationships between test scores and other performance indicators relevant to the broader domain enables evaluation of the extent to which generalizations to the broader domain are supported. This information should be made available to all those who use the test and interpret the test scores.

Standard 13.4

Local norms should be developed when necessary to support test users' intended interpretations.

Comment: Comparison of examinees' scores to local as well as more broadly representative norm groups can be informative. Thus, sample size permitting, local norms are often useful in conjunction with published norms, especially if the local population differs markedly from the population on which published norms are based. In some cases, local norms may be used exclusively.

Standard 13.5

When test results substantially contribute to making decisions about student promotion or graduation, there should be evidence that the test adequately covers only the specific or generalized content and skills that students have had an opportunity to learn.

Comment: Students, parents, and educational staff should be informed of the domains on which the students will be tested, the nature of the item types, and the standards for mastery. Reasonable efforts should be made to document the provision of instruction on tested content and skills, even though it may not be possible or feasible to determine the specific content of instruction for every student. Chapter 7 provides a more thorough discussion of the difficulties that arise with this conception of fairness in testing.

Standard 13.6

Students who must demonstrate mastery of certain skills or knowledge before being promoted or granted a diploma should have a reasonable number of opportunities to succeed on equivalent forms of the test or be provided with construct-equivalent testing alternatives of equal difficulty to demonstrate the skills or knowledge. In most circumstances, when students are provided with multiple opportunities to demonstrate mastery, the time interval between the opportunities should allow for students to have the opportunity to obtain the relevant instructional experiences.

Comment: The number of opportunities and time between each testing opportunity will vary with the specific circumstances of the setting. Further, some students may benefit from a different testing approach to demonstrate their achievement. Care must be taken that evidence of construct equivalence of alternative approaches is provided as well as the equivalence of cut scores defining passing expectations.

Standard 13.7

In educational settings, a decision or characterization that will have major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision.

Comment: As an example, when the purpose of testing is to identify individuals with special needs, including students who would benefit from gifted and talented programs, a screening for eligibility or an initial assessment should be conducted. The screening or initial assessment may in turn call for more comprehensive evaluation. The comprehensive assessment should involve the use of

multiple measures, and data should be collected from multiple sources. Any assessment data used in making decisions are evaluated in terms of validity, reliability, and relevance to the specific needs of the students. It is important that in addition to test scores, other relevant information (e.g., school record, classroom observation, parent report) is taken into account by the professionals making the decision.

Standard 13.8

When an individual student's scores from different tests are compared, any educational decision based on this comparison should take into account the extent of overlap between the two constructs and the reliability or standard error of the difference score.

Comment: When difference scores between two tests are used to aid in making educational decisions, it is important that the two tests are standardized and, if appropriate, normed on the same population at about the same time. In addition, the reliability and standard error of the difference scores between the two tests are affected by the relationship between the constructs measured by the tests as well as the standard errors of measurement of the scores of the two tests. In the case of comparing ability with achievement test scores, the overlapping nature of the two constructs may render the reliability of the difference scores lower than test users normally would assume. If the ability and/or achievement tests involve a significant amount of measurement error, this will also reduce the confidence one may place on the difference scores. All these factors affect the reliability of difference scores between tests and should be considered by professional evaluators in using difference scores as a basis for making important decisions about a student. This standard is also relevant when comparing scores from different components

of the same test such as multiple aptitude test batteries and selection tests.

Standard 13.9

When test scores are intended to be used as part of the process for making decisions for educational placement, promotion, or implementation of prescribed educational plans, empirical evidence documenting the relationship among particular test scores, the instructional programs, and desired student outcomes should be provided. When adequate empirical evidence is not available, users should be cautioned to weigh the test results accordingly in light of other relevant information about the student.

Comment: The validity of test scores for placement or promotion decisions rests, in part, upon evidence about whether students, in fact, benefit from the differential instruction. Similarly, in special education, when test scores are used in the development of specific educational objectives and instructional strategies, evidence is needed to show that the prescribed instruction enhances students' learning. When there is limited evidence about the relationship among test results, instructional plans, and student achievement outcomes, test developers and users should stress the tentative nature of the test-based recommendations and encourage teachers and other decision makers to consider the usefulness of test scores in light of other relevant information about the students.

Standard 13.10

Those responsible for educational testing programs should ensure that the individuals who administer and score the test(s) are proficient in the appropriate test administration procedures and scoring procedures and that they understand the importance of adhering to the directions provided by the test developer.

STANDARDS

EDUCATIONAL TESTING AND ASSESSMENT / PART III

Standard 13.11

In educational settings, test users should ensure that any test preparation activities and materials provided to students will not adversely affect the validity of test score inferences.

Comment: In most educational testing contexts, the goal is to use a sample of test items to make inferences to a broader domain. When inappropriate test preparation activities occur, such as teaching items that are equivalent to those on the test, the validity of test score inferences is adversely affected. The appropriateness of test preparation activities and materials can be evaluated, for example, by determining the extent to which they reflect the specific test items and the extent to which test scores are artificially raised without actually increasing students' level of achievement.

Standard 13.12

In educational settings, those who supervise others in test selection, administration, and interpretation should have received education and training in testing necessary to ensure familiarity with the evidence for validity and reliability for tests used in the educational setting and to be prepared to articulate or to ensure that others articulate a logical explanation of the relationship among the tests used, the purposes they serve, and the interpretations of the test scores.

Standard 13.13

Those responsible for educational testing programs should ensure that the individuals who interpret the test results to make decisions within the school context are qualified to do so or are assisted by and consult with persons who are so qualified.

Comment: When testing programs are used as a strategy for guiding instruction, teachers expected to make inferences about instructional needs may need assistance in interpreting test results for this purpose. If the tests are normed locally, statewide, or nationally, teachers and administrators need to be proficient in interpreting the norm-referenced test scores.

The interpretation of some test scores is sufficiently complex to require that the user have relevant psychological training and experience or be assisted by and consult with persons who have such training and experience. Examples of such tests include individually administered intelligence tests, personality inventories, projective techniques, and neuropsychological tests.

Standard 13.14

In educational settings, score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level and information on how to interpret the scores.

Comment: This information should be communicated in a way that is accessible to persons receiving the score report. For instance, the degree of uncertainty might be indicated by a likely range of scores or by the probability of misclassification.

Standard 13.15

In educational settings, reports of group differences in test scores should be accompanied by relevant contextual information, where possible, to enable meaningful interpretation of these differences. Where appropriate contextual information is not available, users should be cautioned against misinterpretation.

Comment: Observed differences in test scores between groups (e.g., classified by gender, race/ethnicity, school/district, geographical region) can be influenced, for example, by differences in course-taking patterns, in curriculum, in teacher's qualifications, or in parental educational level. Differences in performance of cohorts of students across time may be influenced by changes in the population of students tested or changes in learning opportunities for students. Users should be advised to consider the appropriate contextual information and cautioned against misinterpretation.

Standard 13.16

In educational settings, whenever a test score is reported, the date of test administration should be reported. This information and the age of any norms used for interpretation should be considered by test users in making inferences.

Comment: When a test score is used for a particular purpose, the date of the test score should be taken into consideration in determining its worth or appropriateness for making inferences about a student. Depending on the particular domain measured, the validity of score inferences may be questionable as time progresses. For instance, a reading score from a test administered 6 months ago to an elementary school-aged student may no longer reflect the student's current reading level. Thus, a test score should not be used if it has been determined that undue time has passed since the time of data collection and that the score no longer can be considered a valid indicator of a student's current level of proficiency.

Standard 13.17

When change or gain scores are used, such scores should be defined and their technical qualities should be reported.

Comment: The use of change or gain scores presumes the same test or equivalent forms of the test were used and that the test has (or the forms have) not been materially altered between administrations. The standard error of the difference between scores on the pretest and posttest, the regression of posttest scores on pretest scores, or relevant data from other reliable methods for examining change, such as those based on structural equation modeling, should be reported.

Standard 13.18

Documentation of design, models, scoring algorithms, and methods for scoring and classifying should be provided for tests administered and scored using multimedia or computers. Construct-irrelevant variance pertinent to computer-based testing and the use of other media in testing, such as the test taker's familiarity with technology and the test format, should be addressed in their design and use.

Comment: It is important to assure that the documentation does not jeopardize the security of the items that could adversely affect the validity of score interpretations. Computer and multimedia testing need to be held to the same requirements of technical quality as are other tests.

Standard 13.19

In educational settings, when average or summary scores for groups of students are reported, they should be supplemented with additional information about the sample size and shape or dispersion of score distributions.

Comment: Score reports should be designed to communicate clearly and effectively to their intended audiences. In most cases, reports that go beyond average score comparisons are helpful in furthering thoughtful use

STANDARDS

EDUCATIONAL TESTING AND ASSESSMENT / PART III

and interpretation of test scores. Depending on the intended purpose and audience of the score report, additional information might take the form of standard deviations or other common measures of score variability, or of selected percentile points for each distribution. Alternatively, benchmark score levels might be established and then, for each group or region, the proportions of test takers attaining each specified level could be reported. Such benchmarks might be defined, for example, as selected percentiles of the pooled distribution for all groups or regions. Other distributional summaries of reporting formats may also be useful. The goal of more detailed reporting must be balanced against goals of clarity and conciseness in communicating test scores.

14. TESTING IN EMPLOYMENT AND CREDENTIALING

Background

Employment testing is carried out by organizations for purposes of employee selection, promotion, or placement. *Selection* generally refers to decisions about which individuals will enter the organization; *placement* refers to decisions as to how to assign individuals to positions within the work force; and *promotion* refers to decisions about which individuals within the organization will advance. What all three have in common is a focus on the prediction of future job behaviors, with the goal of influencing organizational outcomes such as efficiency, growth, productivity, and employee motivation and satisfaction.

Testing used in the processes of licensure and certification, which will here generically be called credentialing, focuses on the applicant's current skill or competency in a specified domain. In many occupations, individuals must be licensed by governmental agencies in order to engage in the particular occupation. In other occupations, professional societies or other organizations assume responsibility for credentialing. Although licensure is typically a credential for entry into an occupation, credentialing programs may exist at varying levels, from novice to expert in a given field. Certification is usually sought voluntarily, although occupations differ in the degree to which obtaining certification influences employability or advancement. Testing is commonly only a part of a credentialing process, which may also include other requirements, such as education or supervised experiences. The *Standards* apply to the use of tests in the broader credentialing process.

Testing is also carried out in work organizations for a variety of purposes other than employment decision making and credentialing. Testing to detect psychopathology can take place, as in the case of an employee exhibiting

behavioral problems at work. Testing as a tool for personal growth can be part of training and development programs, in which instruments measuring personality characteristics, interests, values, preferences, and work styles are commonly used with the goal of providing self-insight to employees. Testing can also take place in the context of program evaluation, as in the case of an experimental study of the effectiveness of a training program, where tests may be administered as pre- and post-measures. The focus of this chapter, though, is on the use of testing in employment and credentialing. Many issues relevant to such testing are discussed in other chapters: technical matters in chapters 1-6, fairness issues in chapters 7-10, general issues of test use in chapter 11, and individualized assessment of job candidates in chapter 12.

Employment Testing

THE INFLUENCE OF CONTEXT ON TEST USE

Employment testing involves using test information to aid in personnel decision making. Both the content and the context of employment testing varies widely. Content may cover various domains of knowledge, skills, abilities, traits, dispositions, and values. The context in which tests are used also varies widely. Some contextual features represent choices made by the employing organization; others represent constraints that must be accommodated by the employing organization. Decisions about the design, evaluation, and implementation of a testing system are specific to the context in which the system is to be used. Important contextual features include the following:

Internal vs. external candidate pool.

In some instances, such as promotional settings, the candidates to be tested are already employed by the organization. In others, applications are sought from outside the

organization. In others, a mix of internal and external candidates is sought.

Untrained vs. specialized jobs. In some instances, untrained individuals are selected either because the job does not require specialized knowledge or skill or because the organization plans to offer training after the point of hire. In other instances, trained or experienced workers are sought with the expectation that they can immediately step into a specialized job. Thus, the same job may require very different selection systems depending on whether trained or untrained individuals will be hired or promoted.

Short-term vs. long-term focus. In some instances, the goal of the selection system is to predict performance immediately upon or shortly after hire. In other instances, the concern is with longer-term performance, as in the case of predictions as to whether candidates will successfully complete a multiyear overseas job assignment. Concerns about changing job tasks and job requirements also can lead to a focus on characteristics projected to be necessary for performance on the target job in the future, even if not a part of the job as currently constituted.

Screen in vs. screen out. In some instances, the goal of the selection system is to screen in individuals who will perform well on one set of behavioral or outcome criteria of interest to the organization. In others, the goal is to screen out individuals for whom the risk of pathological, deviant, or criminal behavior on the job is deemed too high. A testing system well suited to one objective may be completely inappropriate for another. That an individual is evaluated as a low risk for engaging in pathological behavior does not imply a prediction that the individual will exhibit high levels of job performance. That a test is predictive of one criterion does not support the inference of linkages to other criteria of interest as well.

Mechanical vs. judgmental decision making. In some instances, test information

is used in a mechanical, standardized fashion. This is the case when scores on a test battery are combined by formula and candidates are selected in strict top-down rank order, or when only candidates above specific cut scores are eligible to continue to subsequent stages of a selection system. In other instances, information from a test is judgmentally integrated with information from other tests and with nontest information to form an overall assessment of the candidate.

Ongoing vs. one-time use of a test. In some instances, a test may be used for an extended period of time in an organization, permitting the accumulation of data and experience about the test in that context. In other instances, concerns about test security are such that repeated use is infeasible, and a new test is required for each test administration. For example, a work-sample test for lifeguards, requiring retrieving a mannequin from the bottom of a pool, is not compromised if candidates possess detailed knowledge of the test in advance. In contrast, a written job knowledge test may be severely compromised if some candidates have access to the test in advance. The key question is whether advance knowledge of test content changes the constructs measured by the test.

Fixed applicant pool vs. continuous flow. In some instances, an applicant pool can be assembled prior to beginning the selection process, as in the case of a policy that all candidates applying before a specific date will be considered. In other cases, there is a continuous flow of applicants about whom employment decisions need to be made on an ongoing basis. A ranking of candidates is possible in the case of the fixed pool; in the case of a continuous flow, a decision may need to be made about each candidate independent of information about other candidates.

Small vs. large sample size. Large sample sizes are sometimes available for jobs with many incumbents, in situations in which multiple similar jobs can be pooled, or in situa-

PART III / TESTING IN EMPLOYMENT AND CREDENTIALING

tions in which organizations with similar jobs collaborate in selection system development. In other situations, sample sizes are small; at the extreme is the case of the single-incumbent job. Sample size affects the degree to which different lines of evidence can be drawn on in examining validity for the intended inference to be drawn from the test. For example, relying on the local setting for empirical linkages between test and criterion scores is not technically feasible with small sample sizes.

Size of applicant pool, relative to the number of job openings. The size of an applicant pool can constrain the type of testing system that is feasible. For desirable jobs, very large numbers of candidates may vie for a small number of jobs. Under such scenarios, short screening tests may be used to reduce the pool to a size for which the administration of more time-consuming and expensive tests is practicable. Large applicant pools may also pose test security concerns, limiting the organization to testing methods that permit simultaneous test administration to all candidates.

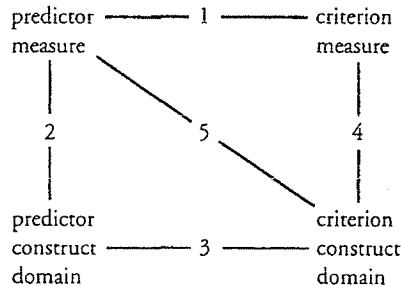
Thus, test use by employers is conditioned by contextual features such as those in the foregoing list. Knowledge of these features plays an important part in the professional judgment that will influence both the type of testing system that will be developed and the strategy that will be used to evaluate critically the validity of the inference(s) drawn using the testing system.

THE VALIDATION PROCESS IN EMPLOYMENT TESTING

The fundamental inference to be drawn from test scores in most applications of testing in employment settings is one of prediction: the test user wishes to make an inference from test results to some future job behavior or job outcome. Even when the validation strategy used does not involve empirical predictor-criterion linkages, as in the case of reliance on validity evidence based on test content, there is an implied criterion. Thus, while different strategies of gathering evidence may be used, the inference to be supported is that scores on

the test can be used to predict subsequent job behavior. The validation process in employment settings involves the gathering and evaluation of evidence relevant to sustaining or challenging this inference. As detailed below, a variety of validation strategies can be used to support this inference.

It thus follows that establishing this predictive inference requires that attention be paid to two domains: that of the test (the predictor) and that of the job behavior or outcome of interest (the criterion). Evaluating the use of a test for an employment decision can be viewed as testing the hypothesis of a linkage between these domains. Operationally, there are many ways of testing this hypothesis. This is illustrated by the following diagram:



The diagram differentiates between a predictor construct domain and a predictor measure and between a criterion construct domain and a criterion measure. A *predictor construct domain* is defined by specifying the set of behaviors that will be included under a particular construct label (e.g., verbal reasoning, typing speed, conscientiousness). Similarly, a *criterion construct domain* specifies the set of job behaviors or job outcomes that will be included under a particular construct label (e.g., performance of core job tasks, teamwork, attendance, sales volume, overall job performance). Predictor and criterion measures are attempts at operationalizing these domains.

TESTING IN EMPLOYMENT AND CREDENTIALING / PART III

The diagram enumerates a number of inferences commonly of interest. The first is the inference that scores on a predictor measure are related to scores on a criterion measure. This inference is tested through empirical examination of relationships between the two measures. The second and fourth are conceptually similar: both examine the inference that an operational measure can be interpreted as representing an individual's standing on the construct domain of interest. Logical analysis, expert judgment, and convergence with or divergence from conceptually similar or different measures are among the forms of evidence that can be examined in testing these linkages. The third is the inference of a relationship between the predictor construct domain and the criterion construct domain. This linkage is established on the basis of theoretical and logical analysis. It commonly draws on systematic evaluation of job content and expert judgment as to the individual characteristics linked to successful job performance. The fifth represents the linkage between the predictor measure and the criterion construct domain.

Some predictor measures are designed explicitly as samples of the criterion construct domain of interest, and, thus, isomorphism between the measure and the construct domain constitutes direct evidence for linkage 5. Establishing linkage 5 in this fashion is the hallmark of approaches that rely heavily on what these *Standards* refer to as "validity evidence based on test content," referred to as content validity in prior conceptualizations of the validation process. Tests in which candidates for life-guard positions perform rescue operations or in which candidates for word processor positions type and edit text exemplify this approach.

A prerequisite to the use of a predictor measure for personnel selection is that the linkage between the predictor measure and the criterion construct domain be established. As the diagram illustrates, there are multiple strategies for establishing this crucial linkage. One strategy is direct, via linkage 5; a second

involves pairing linkage 1 and linkage 4; and a third involves pairing linkage 2 and linkage 3.

When the test is designed as a sample of the criterion construct domain, this linkage can be established directly via linkage 5. Another strategy for linking a predictor measure and the criterion construct domain focuses on linkages 1 and 4: pairing an empirical link between the predictor and criterion measures with evidence of the adequacy with which the criterion measure represents the criterion construct domain. The empirical link between the predictor measure and the criterion measure is part of what these *Standards* refer to as "validity evidence based on relationships to other variables," referred to as criterion-related validity in prior conceptualizations of the validation process. The empirical link of the test and the criterion measure must be supplemented by evidence of the relevance of the criterion measure to the criterion construct domain to complete the linkage between the test and the criterion construct domain. Evidence of the relevance of the criterion measure to the criterion construct domain is commonly based on job analysis, though in some cases the link between the domain and the measure is so direct that relevance is apparent without job analysis (e.g., when the criterion construct of interest is absenteeism or turnover). Note that this strategy does not necessarily rely on a well-developed predictor construct domain. Predictor measures such as empirically keyed biodata measures are constructed on the basis of empirical links between test item responses and the criterion measure of interest. Such measures may, in some instances, be developed without a fully established a priori conception of the predictor construct domain; the basis for their use is the direct empirical link between test responses and a relevant criterion measure.

Yet another strategy for linking predictor scores and the criterion construct domain focuses on pairing evidence of the adequacy with which the predictor measure represents the predictor construct domain (linkage 2)

PART III / TESTING IN EMPLOYMENT AND CREDENTIALING

with evidence of the linkage between the predictor construct domain and the criterion construct domain (linkage 3). As noted above, there is no single direct route to establishing these linkages. They involve lines of evidence subsumed under "construct validity" in prior conceptualizations of the validation process. A combination of lines of evidence, such as expert judgment of the characteristics predictive of job success, inferences drawn from an analysis of critical incidents of effective and ineffective job performance, and interview and observation methods, may support inferences about the predictor constructs linked to the criterion construct domain. Measures of these predictor constructs may then be selected or developed, and the linkage between the predictor measure and the predictor construct domain can be established with various lines of evidence for linkage 2 discussed above.

Thus multiple sources of data and multiple lines of evidence can be drawn on to evaluate the linkage between a predictor measure and the criterion construct domain of interest. There is not a single correct or even a preferred method of inquiry for establishing this linkage. Rather, the test user must consider the specifics of the testing situation and apply professional judgment in developing a strategy for testing the hypothesis of a linkage between the predictor measure and the criterion domain.

For many testing applications, there is a considerable cumulative body of research that speaks to some, if not all, of the inferences discussed above. A meta-analytic integration of this research can form an integral part of the strategy for linking test information to the construct domain of interest. The value of collecting local validation data varies with the magnitude, relevance, and consistency of research findings using similar predictor measures and similar criterion construct domains for similar jobs. In some cases, a small and inconsistent cumulative research record may lead to a validation strategy that relies heavily on local data; in others, a large, consistent

research base may make investing resources in additional local data collection unnecessary.

BASES FOR EVALUATING TEST USE

While a primary goal of employment testing is the accurate prediction of subsequent job behaviors or job outcomes, it is important to recognize that there are limits to the degree to which such criteria can be predicted. Perfect prediction is an unattainable goal. First, behavior in work settings is also influenced by a wide variety of organizational and extra-organizational factors, including supervisor and peer coaching, formal and informal training, changes in job design, changes in organizational structures and systems, and changing family responsibilities, among others. Second, behavior in work settings is influenced by a wide variety of individual characteristics, including knowledge, skills, abilities, personality, and work attitudes, among others. Thus any single characteristic will be only an imperfect predictor, and even complex selection systems focus on the set of constructs deemed most critical for the job, rather than on all characteristics that can influence job behavior. Third, some measurement error always occurs even in well-developed test and criterion measures.

Thus, testing systems cannot be judged against a standard of perfect prediction but rather in terms of comparisons with available alternative selection methods. Professional judgment, informed by knowledge of the research literature about the degree of predictive accuracy relative to available alternatives, influences decisions about test use.

Decisions about test use are often influenced by additional considerations including utility (i.e., cost-benefit) evaluation, value judgments about the relative importance of selecting for one criterion domain vs. others, concerns about applicant reactions to test content and process, the availability and appropriateness of alternative selection methods, statutory or regulatory requirements governing test use, and social issues such as workforce

diversity. Organizational values necessarily come into play in making decisions about test use; organizations with comparable evidence supporting an intended inference drawn from test scores may thus reach different conclusions about whether to use any particular test.

Testing in Professional and Occupational Credentialing

Tests are widely used in the credentialing of persons for many occupations and professions. Licensing requirements are imposed by state and local governments to ensure that those licensed possess knowledge and skills in sufficient degree to perform important occupational activities safely and effectively. Certification plays a similar role in many occupations not regulated by governments and is often a necessary precursor to advancement in many occupations. Certification has also become widely used to indicate that a person has certain specific skills (e.g., operation of specialized auto repair equipment) or knowledge (e.g., estate planning), which may be only a part of their occupational duties. Licensure and certification, as well as registry and other warrants of expertise, will here generically be called credentialing.

Tests used in credentialing are intended to provide the public, including employers and government agencies, with a dependable mechanism for identifying practitioners who have met particular standards. The standards are strict, but not so stringent as to unduly restrain the right of qualified individuals to offer their services to the public. Credentialing also serves to protect the profession by excluding persons who are deemed to be not qualified to do the work of the occupation. Qualifications for credentials typically include educational requirements, some amount of supervised experience, and other specific criteria, as well as attainment of a passing score on one or more examinations. Tests are used in credentialing in a broad spectrum of profes-

sions and occupations, including medicine, law, psychology, teaching, architecture, real estate, and cosmetology. In some of these, such as actuarial science, clinical neuropsychology, and medical specialties, tests are also used to certify advanced levels of expertise. Relicensure or recertification is also required in some occupations and professions.

Tests used in credentialing are designed to determine whether the essential knowledge and skills of a specified domain have been mastered by the candidate. The focus of performance standards is on levels of knowledge and performance necessary for safe and appropriate practice. Test design generally starts with an adequate definition of the occupation or specialty, so that persons can be clearly identified as engaging in the activity. Then, the nature and requirements of the occupation, in its current form, are delineated. Often, a thorough analysis is conducted of the work performed by people in the profession or occupation to document the tasks and abilities that are essential to practice. A wide variety of empirical approaches is used, including delineation, critical incidence techniques, job analysis, training needs assessments, or practice studies and surveys of practicing professionals. Panels of respected experts in the field often work in collaboration with qualified specialists in testing to define test specifications, including the knowledge and skills needed for safe, effective performance, and an appropriate way of assessing that performance. Forms of testing may include traditional multiple-choice tests, written essays, and oral examinations. More elaborate performance tasks, sometimes using computer-based simulation, are also used in assessing such practice components as, for example, patient diagnosis or treatment planning. Hands-on performance tasks may also be used (e.g., operating a boom crane or filling a tooth) while being observed by one or more examiners.

Credentialing tests may cover a number of related but distinct areas. Designing the testing

PART III / TESTING IN EMPLOYMENT AND CREDENTIALING

program includes deciding what areas are to be covered, whether one or a series of tests is to be used, and how multiple test scores are to be combined to reach an overall decision. In some cases high scores on some tests are permitted to offset low scores on other tests, so that additive combination is appropriate. In other cases, an acceptable performance level is required on each test in an examination series.

Validation of credentialing tests depends mainly on content-related evidence, often in the form of judgments that the test adequately represents the content domain of the occupation or specialty being considered. Such evidence may be supplemented with other forms of evidence external to the test. Criterion-related evidence is of limited applicability in licensure settings because criterion measures are generally not available for those who are not granted a license.

Defining the minimum level of knowledge and skill required for licensure or certification is one of the most important and difficult tasks facing those responsible for credentialing. Verifying the appropriateness of the cut score or scores on the tests is a critical element in validity. The validity of the inference drawn from the test depends on whether the standard for passing makes a valid distinction between adequate and inadequate performance. Often, panels of experts are used to specify the level of performance that should be required. Standards must be high enough to protect the public, as well as the practitioner, but not so high as to be unreasonably limiting. Verifying the appropriateness of the cut score or scores on a test used for licensure or certification is a critical element of the validity of test results.

Legislative bodies sometimes attempt to legislate a cut score, such as a score of 70%. Arbitrary numerical specifications of cut scores are unhelpful for two reasons. First, without detailed information about the test, job requirements, and their relationship, sound standard setting is impossible. Second, without

detailed information about the format of the test and the difficulty of items, such numerical specifications have little meaning.

Tests for credentialing need to be precise in the vicinity of the passing, or cut, score. They may not need to be precise for those who clearly pass or clearly fail. Sometimes a test used in credentialing is designed to be precise only in the vicinity of the cut score. Computer-based mastery tests may include a procedure to end the testing when a decision about the candidate's performance can be clearly made or when a maximum time limit is reached. This may result in a shorter test for candidates whose performance clearly exceeds or falls far below the minimum performance required for a passing score. The test taker may be told only whether the decision was pass or fail. Because such mastery tests are not designed to indicate how badly the candidate failed, or how well the candidate passed, providing scores that are much higher or lower than the cut score could be misleading. Nevertheless, candidates who fail are likely to profit from information about the areas in which their performance was especially weak. When feedback to candidates about how well or how poorly they performed is intended, precision throughout the score range is needed.

Practice in professions and occupations often changes over time. Evolving legal restrictions, progress in scientific fields, and refinements in techniques can result in a need for changes in test content. When change is substantial, it becomes necessary to revise the definition of the job, and the test content, to reflect changing circumstances. When major revisions are made in the test, the cut score that identifies required test performance is also reestablished.

Because credentialing is an ongoing process, with tests given on a regular schedule, new versions of the test are often needed. From a technical perspective, all versions of a test should be prepared to the same specifications and represent the same content.

STANDARDS

TESTING IN EMPLOYMENT AND CREDENTIALING / PART III

Alternate test forms should have comparable score scales so that scores can retain their meaning. Various methods of jointly calibrating alternate forms can be used to assure that the standard for passing represents the same level of performance on all forms. It may be noted that release of past test forms may compromise the quality of test form comparability.

Some credentialing groups consider it necessary, as a practical matter, to adjust their criteria yearly in order to regulate the number of accredited candidates entering the profession. This questionable procedure raises serious problems for the technical quality of the test scores. Adjusting the cut score annually implies higher standards in some years than in others, which, although open and straightforward, is difficult to justify on the grounds of quality of performance. Adjusting the score scale so that a certain number or proportion reach the passing score, while less obvious to the candidates, is technically inappropriate because it changes the meaning of the scores from year to year. Passing a credentialing examination should signify that the candidate meets the knowledge and skill standards set by the credentialing body, independent of the availability of work.

Issues of cheating and test security are of special importance for testing practices in credentialing. Issues of test security are covered in chapters 5 and 11. Issues of cheating by test takers are covered in chapter 8. Issues concerning the technical quality of tests are found in chapters 1-6, and issues of fairness in chapters 7-10.

Standard 14.1

Prior to development and implementation of an employment test, a clear statement of the objective of testing should be made. The subsequent validation effort should be designed to determine how well the objective has been achieved.

Comment: The objectives of employment tests can vary considerably. Some aim to screen out those least suited for the job in question, while others are designed to identify those best suited for the job. Tests also vary in the aspects of job behavior they are intended to predict, which may include quantity or quality of work output, tenure, counterproductive behavior, and teamwork, among others.

Standard 14.2

When a test is used to predict a criterion, the decision to conduct local empirical studies of predictor-criterion relationships and interpretation of the results of local studies of predictor-criterion relationships should be grounded in knowledge of relevant research.

Comment: The cumulative literature on the relationship between a particular type of predictor and type of criterion may be sufficiently large and consistent to support the predictor-criterion relationship without additional research. In some settings, the cumulative research literature may be so substantial and so consistent that a dissimilar finding in a local study should be viewed with caution unless the local study is exceptionally sound. Local studies are of greatest value in settings where the cumulative research literature is sparse (e.g., due to the novelty of the predictor and/or criterion used), where the cumulative record is inconsistent, or where the cumulative literature does not include studies similar to the local setting (e.g., a test with a

large cumulative literature dealing exclusively with production jobs, and a local setting involving managerial jobs).

Standard 14.3

Reliance on local evidence of empirically determined predictor-criterion relationships as a validation strategy is contingent on a determination of technical feasibility.

Comment: Meaningful evidence of predictor-criterion relationships is conditional on a number of features, including (a) the job being relatively stable, rather than in a period of rapid evolution; (b) the availability of a relevant and reliable criterion measure; (c) the availability of a sample reasonably representative of the population of interest; and (d) an adequate sample size for estimating the strength of the predictor-criterion relationship.

Standard 14.4

When empirical evidence of predictor-criterion relationships is part of the pattern of evidence used to support test use, the criterion measure(s) used should reflect the criterion construct domain of interest to the organization. All criteria used should represent important work behaviors or work outputs, on the job or in job-relevant training, as indicated by an appropriate review of information about the job.

Comment: When criteria are constructed to represent job activities or behaviors (e.g., supervisory ratings of subordinates on important job dimensions), systematic collection of information about the job informs the development of the criterion measures, though there is no clear choice among the many available job analysis methods. There is not a clear need for job analysis to support criterion use when measures such as absenteeism or turnover are the criteria of interest.

Standard 14.5

Individuals conducting and interpreting empirical studies of predictor-criterion relationships should identify contaminants and artifacts that may have influenced study findings, such as error of measurement, range restriction, and the effects of missing data. Evidence of the presence or absence of such features, and of actions taken to remove or control their influence, should be retained and made available as needed.

Comment: Error of measurement in the criterion and restriction in the variability of predictor or criterion scores systematically reduce estimates of the relationship between predictor measures and the criterion construct domain, and procedures for correction for the effects of these artifacts are available. When these procedures are applied, both corrected and uncorrected values should be presented, along with the rationale for the correction procedures chosen. Statistical significance tests for uncorrected correlations should not be used with corrected correlations. Other features to be considered include issues such as missing data for some variables for some individuals, decisions about the retention or removal of extreme data points, the effects of capitalization on chance in selecting predictors from a larger set on the basis of strength of predictor-criterion relationships, and the possibility of spurious predictor-criterion relationships, as in the case of collecting criterion ratings from supervisors who know selection test scores.

Standard 14.6

Evidence of predictor-criterion relationships in a current local situation should not be inferred from a single previous validation study unless the previous study of the predictor-criterion relationship was done under favorable conditions (i.e., with a large sample size and a relevant criterion) and if the current situation corresponds closely to the previous situation.

STANDARDS

TESTING IN EMPLOYMENT AND CREDENTIALING / PART III

Comment: Close correspondence means that the job requirements or underlying psychological constructs are substantially the same (as is determined by a job analysis), and that the predictor is substantially the same.

Standard 14.7

If tests are to be used to make job classification decisions (e.g., the pattern of predictor scores will be used to make differential job assignments), evidence that scores are linked to different levels or likelihoods of success among jobs or job groups is needed.

Standard 14.8

Evidence of validity based on test content requires a thorough and explicit definition of the content domain of interest. For selection, classification, and promotion, the characterization of the domain should be based on job analysis.

Comment: In general, the job content domain should be described in terms of job tasks or worker knowledge, skills, abilities, and other personal characteristics that are clearly operationally defined so that they can be linked to test content, and for which job demands are not expected to change substantially over a specified period of time. Knowledge, skills, and abilities included in the content domain should be those the applicant should already possess when being considered for the job in question.

Standard 14.9

When evidence of validity based on test content is a primary source of validity evidence in support of the use of a test in selection or promotion, a close link between test content and job content should be demonstrated.

Comment: For example, if the test content samples job tasks with considerable fidelity

(e.g., actual job samples such as machine operation) or, in the judgment of experts, correctly simulates job task content (e.g., certain assessment center exercises), or samples specific job knowledge required for successful job performance (e.g., information necessary to exhibit certain skills), then content-related evidence can be offered as the principal form of evidence of validity. If the link between the test content and the job content is not clear and direct, other lines of validity evidence take on greater importance.

Standard 14.10

When evidence of validity based on test content is presented, the rationale for defining and describing a specific job content domain in a particular way (e.g., in terms of tasks to be performed or knowledge, skills, abilities, or other personal characteristics) should be stated clearly.

Comment: When evidence of validity based on test content is presented for a job or class of jobs, the evidence should include a description of the major job characteristics that a test is meant to sample, including the relative frequency, importance, or criticality of the elements.

Standard 14.11

If evidence based on test content is a primary source of validity evidence supporting the use of a test for selection into a particular job, a similar inference should be made about the test in a new situation only if the critical job content factors are substantially the same (as is determined by a job analysis), the reading level of the test material does not exceed that appropriate for the new job, and there are no discernible features of the new situation that would substantially change the original meaning of the test material.

Standard 14.12

When the use of a given test for personnel selection relies on relationships between a predictor construct domain that the test represents and a criterion construct domain, two links need to be established. First, there should be evidence for the relationship between the test and the predictor construct domain, and second, there should be evidence for the relationship between the predictor construct domain and major factors of the criterion construct domain.

Comment: There should be a clear conceptual rationale for these linkages. Both the predictor construct domain and the criterion construct domain to which it is to be linked should be defined carefully. There is no single route to establishing these linkages. Evidence in support of linkages between the two construct domains can include patterns of findings in the research literature and systematic evaluation of job content to identify predictor constructs linked to the criterion domain. The bases for judgments linking the predictor and criterion construct domains should be articulated.

Standard 14.13

When decision makers integrate information from multiple tests or integrate test and nontest information, the role played by each test in the decision process should be clearly explicated, and the use of each test or test composite should be supported by validity evidence.

Comment: A decision maker may integrate test scores with interview data, reference checks, and many other sources of information in making employment decisions. The inferences drawn from test scores should be limited to those for which validity evidence is available. For example, viewing a high test score as indicating overall job suitability, and

thus precluding the need for reference checks, would be an inappropriate inference from a test measuring a single narrow, albeit relevant, domain, such as job knowledge. In other circumstances, decision makers integrate scores across multiple tests, or across multiple scales within a given test.

Standard 14.14

The content domain to be covered by a credentialing test should be defined clearly and justified in terms of the importance of the content for credential-worthy performance in an occupation or profession. A rationale should be provided to support a claim that the knowledge or skills being assessed are required for credential-worthy performance in an occupation and are consistent with the purpose for which the licensing or certification program was instituted.

Comment: Some form of job or practice analysis provides the primary basis for defining the content domain. If the same examination is used in the licensure or certification of people employed in a variety of settings and specialties, a number of different job settings may need to be analyzed. Although the job analysis techniques may be similar to those used in employment testing, the emphasis for licensure is limited appropriately to knowledge and skills necessary for effective practice. The knowledge and skills contained in a core curriculum designed to train people for the job or occupation may be relevant, especially if the curriculum has been designed to be consistent with empirical job or practice analyses. In tests used for licensure, skills that may be important to success but are not directly related to the purpose of licensure (e.g., protecting the public) should not be included. For example, in real estate, marketing skills may be important for success as a broker, and assessment of these skills might have utility for agencies selecting brokers for

STANDARDS

TESTING IN EMPLOYMENT AND CREDENTIALING / PART III

employment. However, lack of these skills may not present a threat to the public and would appropriately be excluded from consideration for a licensing examination. The fact that successful practitioners possess certain knowledge or skills is relevant but not persuasive. Such information needs to be coupled with an analysis of the purpose of a licensing program and the reasons that the knowledge or skill is required in an occupation or profession.

Standard 14.15

Estimates of the reliability of test-based credentialing decisions should be provided.

Comment: The standards for decision reliability described in chapter 2 are applicable to tests used for licensure and certification. Other types of reliability estimates and associated standard errors of measurement may also be useful, but the reliability of the decision of whether or not to certify is of primary importance.

Standard 14.16

Rules and procedures used to combine scores on multiple assessments to determine the overall outcome of a credentialing test should be reported to test takers, preferably before the test is administered.

Comment: In some cases, candidates may be required to score above a specified minimum on each of several tests. In other cases, the pass-fail decision may be based solely on a total composite score. While candidates may be told that tests will be combined into a composite, the specific weights given to various components may not be known in advance (e.g., to achieve equal effective weights, nominal weights will depend on the variance of the components).

Standard 14.17

The level of performance required for passing a credentialing test should depend on the knowledge and skills necessary for acceptable performance in the occupation or profession and should not be adjusted to regulate the number or proportion of persons passing the test.

Comment: The number or proportion of persons granted credentials should be adjusted, if necessary, on some basis other than modifications to either the passing score or the passing level. The cut score should be determined by a careful analysis and judgment of acceptable performance. When there are alternate forms of the test, the cut score should be carefully equated so that it has the same meaning for all forms.

15. TESTING IN PROGRAM EVALUATION AND PUBLIC POLICY

Background

Tests are widely used in program evaluation and in public policy decision making. Program evaluation is the set of procedures used to make judgments about the client's need for a program, the way it is implemented, its effectiveness, and its value. Policy studies are somewhat broader than program evaluations and refer to studies that contribute to judgments about plans, principles, or procedures enacted to achieve broad public goals. There is no sharp distinction between policy studies and program evaluations, and in many instances there is substantial overlap between the two types of investigations. Test results are often one important source of evidence for the initiation, continuation, modification, termination, or expansion of various programs and policies.

Interpretation of test scores in program evaluation and policy studies usually entails the complex analysis of a number of variables. For example, some programs are mandated for a broad population; others target only certain subgroups. Some are designed to affect attitudes, while others are intended to have a more direct impact on behavior. It is important that the participants included in any study at least meet the specified criteria for the program or policy under review so that appropriate interpretation of test results will be possible. Test results will reflect not only the effects of rules for participant selection and the impact of participation in different programs or treatments, but also the characteristics of those tested. Relevant background information about clients or students may be obtained in order to strengthen the inferences derived from the test results. Valid interpretations may depend upon additional considerations that have nothing to do with the appropriateness of the test or its technical quality, including study design, administrative feasibility, and the quality of

other available data. It is not the intent of this chapter to deal with these varied considerations in any substantial way. In order to develop defensible conclusions, however, investigators conducting program evaluations and policy studies are encouraged to supplement test results with data from other sources. These include information about program characteristics, delivery, costs, client backgrounds, degree of participation, and evidence of side effects. Because test results lend important weight to evaluation and policy studies, it is critical that any tests used in these investigations be sensitive to the questions of the study and appropriate for the test takers.

It is important to evaluate any proposed test in terms of its relevance to the goals of the program or policy and/or to the particular question its use will address. It is relatively rare for a test to be designed specifically for program evaluation or policy study purposes. Typically, the instruments used in such studies were originally developed for purposes other than program or policy evaluation. In addition, because of cost or convenience, certain tests may be adopted for use in a program evaluation or policy study even though they may have been developed for a somewhat different population of respondents. Some tests may be selected for use in program evaluation or policy studies because the tests are well known and thought to be especially credible to the clients or the public consumer. Even though certain tests may be more familiar to the public or may be less time-consuming or less expensive to use than an instrument developed specifically for the evaluation, they may be nonetheless inappropriate for use as criterion measures to determine the need for or to evaluate the effects of particular interventions.

As government agencies and other institutions move to improve their own routine data collection capability, fewer special studies are

conducted to evaluate programs and policies. Instead, evaluations and policy studies may depend upon a special analysis of data previously collected for other purposes. In these cases, the investigators may reanalyze test data already obtained and analyzed for another purpose in order to make inferences about program or policy effectiveness. This procedure is called *secondary data analysis*. In some circumstances, it may be difficult to assure a good match between the existing test and the intervention or the policy under examination. Moreover, it may be difficult to reconstruct in detail the conditions under which the data were originally collected. Secondary data analysis also requires consideration of whether adequate informed consent was obtained from subjects in the original data collection to allow secondary analysis to occur without obtaining additional consent. In selecting (or developing) a test or in deciding to use existing data in evaluation and policy studies, careful investigators attempt to balance the purpose of the test, its likelihood to be sensitive to the intervention under study, the credibility of the test to interested parties, and the costs of its administration. Otherwise, test results may lead to inappropriate interpretations about the progress, impact, and overall value of programs and policies under review.

Program Evaluation

Tests may be used in program evaluations to provide information on the status of clients or students before, during, or following an intervention, as well as to provide information on appropriate comparison groups. Whereas understanding the performance of an individual student or client is often the goal of many testing activities, program evaluation targets the performance of, or impact on, groups. Tests are used in program evaluations in a variety of fields, such as social services, education, health services, and military and employment training. The term *program*, broadly interpret-

ed, describes interventions that range from large-scale state or national programs with provisions for local flexibility to small-scale, more experimental projects. In many cases, evaluation is mandated by the agency or funding source for the program, and the intervention is evaluated by judging its effectiveness in meeting stated goals. Some examples of programs that might use test results as part of their evaluation data include psychotherapeutic services, military training programs and job placement programs, school curricula, or services for individuals with special needs.

Test results, along with other information, may be used to compare competing interventions, such as alternative reading curricula or different psychotherapeutic interventions, or to describe the long-term pattern of effects for one or more groups. It is often important to assess a program for its differential effectiveness in meeting the needs of subgroups (such as different ethnic or gender groups within the target population). Even though the performance of groups is of primary interest in program evaluation, the analysis of individuals' histories and test performances may provide additional useful information to aid in the interpretation of test results.

Because of administrative realities, such as cost constraints and response burden, methodological refinements may be adopted to increase the efficiency of testing. One strategy is to obtain a sample of participants to be evaluated from the larger set of those exposed to a program or policy. When there is a sufficient number of clients affected by the program or policy to be evaluated, and when there is a desire to limit the time spent on testing, evaluators can create multiple forms of shorter tests from a larger pool of items. By constructing a number of different test forms consisting of relatively few items and assigning these test forms to different subsamples of test takers (a procedure known as matrix sampling), a larger number of items can be included in the study than could reasonably be administered to any

PART III / TESTING IN PROGRAM EVALUATION AND PUBLIC POLICY

single test taker. When it is desirable to represent a domain with a large number of test items, this approach is often used. However, individual scores are not usually created or interpreted when matrix sampling is employed. Because procedures for sampling individuals or test items may vary in a number of ways, adequate analysis and interpretation of test results for any study depend upon a clear description of how samples were formed and the manner in which test results were aggregated.

Policy Uses of Tests

As noted previously, tests are also used in policy analyses, and the distinction between program evaluation and policy uses of tests is often a matter of degree. Programs are expected to share particular goals, procedures, and resources. Policy is a broader term, applying to plans, principles, procedures, or programs enacted to achieve particular goals in different settings. Programs provide direct services or interventions. Policies may be constructed to achieve their goals by direct or indirect means. Indeed, one direct approach used to achieve a policy goal might include the funding of specific programs. Other examples of direct policy approaches might involve the provision of training resources to improve performance in particular health-service occupations, or the enactment of new recertification requirements for accountants. Studies of the need for or impact of both of these policies could in part depend upon the analyses of test results. To illustrate in more depth, to meet the general policy objective of containing the costs of health care, direct policies might include giving incentives to clients to participate in fitness programs and the development of patient education programs. Tests could measure the understandings and attitudes of participants about the relationship of fitness to the prevention of illness. Another policy example, using a more indirect approach, is to encourage educators to create more effective programs for

children from low-income families. As an approach, a state's educational authorities might require the separate reporting of test scores for children in high-poverty areas. Large differences in group performance would be expected to attract the attention of the public and to place greater pressure on the schools to improve the performance of particular groups of children.

In decentralized governments, policy implementation may be left to local authorities and may be interpreted in a number of different ways. As a result, it may be difficult to select or develop a single test or outcome measure that will be sensitive to the range of different activities or tactics used to implement a given policy. For that reason, policy studies may often use more than one test or outcome measure to provide a more adequate picture of the range of effects.

Issues in Program and Policy Evaluation

Test results are sometimes used as one way to inspire program administrators as well as to infer institutional effectiveness. This use of tests, including the public reporting of results, is thought to encourage an institution to improve its services for its clients. For example, consistently poor achievement test results may trigger special management attention for public schools in some locales. The interpretation of test results is especially complex when tests are used both as an institutional policy mechanism and as a measure of effectiveness. For example, a policy or program may be based on the assumption that providing clear goals and general specifications of test content (such as the type of topics, constructs and cognitive domains, and responses included in the test) may be a reasonable strategy to communicate new expectations to educators. Yet, the desire to influence test or evaluation results to show acceptable institutional performance could lead to inappropriate testing practices, such as

TESTING IN PROGRAM EVALUATION AND PUBLIC POLICY / PART III

teaching the test items in advance, modifying test administration procedures, discouraging certain students or clients from participating in the testing sessions, or focusing exclusively on test-taking procedures. These practices might occur instead of those aimed at helping the test taker learn the domains measured by the test. Because results derived from such practices might lead to spuriously high estimates of impact and might reflect the negative side effects of this particular policy, diligent investigators may estimate the impact of such consequences in order to interpret the test results appropriately. Looking at possible inappropriate consequences of tests as well as their benefits will better assess policy claims that particular types of testing programs lead to improved performance.

On the other hand, policy studies and program evaluations often do not make available reports of results to the test takers and may give no clear reasons to the test taker for participating in the testing procedure. For example, when matrix sampling is used for program evaluation, it may not be feasible to provide such reports. If little effort is made to motivate the test taker to regard the test seriously (for instance, if the purpose of the test is not explained to the test taker), it is possible that test takers might have little reason to try to perform well on the test. Obtained test results then might well underrepresent the impact of the program, institution, or policy because of poor motivation on the part of the test taker. When there is a suspicion that the test might not have been taken seriously, motivation of test takers may be explored by collecting additional information, using observation or interview methods. The issues of inappropriate preparation or unmotivated performance are examples that raise basic questions about the validity of interpretations of test results. In every case, it is important to consider the potential impact of the testing process itself, including test administration and reporting practices, on the test taker.

Public policy decisions are rarely based solely on the results of empirical studies, even when the studies have been well done. The more expansive and indirect the policy, the more likely will it be that other considerations will come into play, such as the political and economic impact of abandoning, changing, or retaining the policy, or the reaction to offering rewards or sanctions to institutions. In a political climate, tests used in policy settings may be subjected to intense and detailed scrutiny. When results do not support a favored position, attempts may be made to discount the appropriateness of the testing procedure, construct, or interpretation.

It is important that all tests used in public evaluation or policy contexts meet the standards described in earlier chapters. As described in chapter 8, tests are to be administered by trained personnel. It is also essential that assistance be provided to those responsible for interpreting study results to practitioners, to the lay public, and to the media. Careful communication of the study's goals, procedures, findings, and limitations increases the chances that the public's interpretations will be accurate and useful.

Additional Considerations

This chapter and its associated standards are directed to users of tests in program evaluation and policy studies and to the conditions under which those studies are usually conducted. Other standards documents that are relevant to this chapter include *The Program Evaluation Standards: How to Assess Evaluations of Educational Programs*, prepared by the Joint Committee on Standards for Educational Evaluation (2nd ed., Thousand Oaks, CA: Sage Publications, 1994), and the *Code of Fair Testing Practices in Education*, prepared by the Joint Committee on Testing Practices (Washington, DC: Joint Committee on Testing Practices, 1988).

STANDARDS

Standard 15.1

When the same test is designed or used to serve multiple purposes, evidence of technical quality for each purpose should be provided.

Comment: In educational testing, for example, it has become common practice to use the same test for multiple purposes (e.g., monitoring achievement of individual students, providing information to assist in instructional planning for individuals or groups of students, evaluating schools or districts). No test will serve all purposes equally well. Choices in test development and evaluation that enhance validity for one purpose may diminish validity for other purposes. Different purposes require somewhat different kinds of technical evidence, and appropriate evidence of technical quality for each purpose should be provided by the test developer. If the test user wishes to use the test for a purpose not supported by the available evidence, it is incumbent on the user to provide the necessary additional evidence.

Standard 15.2

Evidence should be provided of the suitability of a test for use in evaluation or policy studies, including the relevance of the test to the goals of the program or policy under study and the suitability of the test for the populations involved.

Comment: Faulty inferences may be made when test scores are not sensitive to the features of a particular intervention. For instance, a test designed for selection may be ineffective as a measure of the effects of an intervention. It is also important to employ tests that are appropriate for the age and background of test takers.

Standard 15.3

When change or gain scores are used, the definition of such scores should be made explicit, and their technical qualities should be reported.

Comment: The use of change or gain scores presumes that the same test or equivalent forms of the test were used and that the test (or forms) have not been materially altered between administrations. The standard error of the difference between scores on pretests and posttests, the regression of posttest scores on pretest scores, or relevant data from other reliable methods for examining change, such as those based on structural equation modeling, should be reported.

Standard 15.4

In program evaluation or policy studies, investigators should complement test results with information from other sources to generate defensible conclusions based on the interpretation of test results.

Comment: Descriptions or analyses of such variables as client selection criteria, services, clients, setting, and resources are often needed to provide a comprehensive picture of the program or policy under review and to aid in the interpretation of test results. Performance on indicators other than tests is almost always useful and in many cases is essential. Examples of other information include attrition rates or patterns of participation. Another source of information might be to determine the degree of motivation of the test takers. When individual scores are not reported to test takers, it is important to determine whether the examinees took the test experience seriously.

STANDARDS

TESTING IN PROGRAM EVALUATION AND PUBLIC POLICY / PART III

Standard 15.5

Agencies using tests to conduct program evaluations or policy studies, or to monitor outcomes, should clearly describe the population the program or policy is intended to serve and should document the extent to which the sample of test takers is representative of that population.

Comment: For example, a clinic with a diverse client population using testing to assess the outcome of a particular treatment may routinely report the extent of participation by subgroups of clients, for instance, those of diverse ethnic backgrounds or for whom English is a second language.

Standard 15.6

When matrix sampling procedures are used for program evaluation or population descriptions, rules for sampling items and test takers should be provided, and reliability analyses must take the sampling scheme into account.

Standard 15.7

When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to identify and monitor their impact and to minimize potential negative consequences. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test user.

Comment: Mandated testing programs are often justified in terms of their potential benefits for teaching and learning. Concerns have been raised about the potential negative impact of mandated testing programs, particularly when they affect important deci-

sions for individuals or institutions. To the extent possible, students, parents, and staff should be informed of the domains on which the students will be tested, the nature of the item types, and the standards for mastery. Effort should be made to document the provision of instruction in tested content and skills, even though it may not be possible or feasible to determine the specific content of instruction for every student. An example of negative impact is the use of strategies to raise performance artificially.

Standard 15.8

When it is clearly stated or implied that a recommended test use will result in a specific outcome, the basis for expecting that outcome should be presented, together with relevant evidence.

Comment: A given claim for the benefits of test use, such as improving students' achievement, may be supported by logical or theoretical argument as well as empirical data. Due weight should be given to findings in the scientific literature that may be inconsistent with the stated claim.

Standard 15.9

The integrity of test results should be maintained by eliminating practices designed to raise test scores without improving performance on the construct or domain measured by the test.

Comment: Such practices may include teaching test items in advance, modifying test administration procedures, and discouraging or excluding certain test takers from taking the test. These practices can lead to spuriously high scores that do not reflect performance on the underlying construct or domain of interest.

Standard 15.10

Those who have a legitimate interest in an assessment should be informed about the purposes of testing, how tests will be administered and scored, how long records will be retained, and to whom and under what conditions the records may be released.

Comment: Those with a legitimate interest may include the test takers, their parents or guardians, or personnel who may be affected by results (teachers, program staff).

Standard 15.11

When test results are released to the public or to policymakers, those responsible for the release should provide and explain any supplemental information that will minimize possible misinterpretations of the data.

Comment: The context and limitations of the study should be described, with particular attention given to methods of causal inferences.

Standard 15.12

Reports of group differences in average test scores should be accompanied by relevant contextual information, where possible, to enable meaningful interpretation of these differences. Where appropriate contextual information is not available, users should be cautioned against misinterpretation.

Comment: Observed differences in average test scores between groups (e.g., classified by gender, race/ethnicity, or geographical region) can be influenced, for example, by differences in life experiences, training experience, effort, instructor quality, or level and type of parental support. In education, differences in group performance across time may be influenced by changes in the population of those tested or changes in their experiences. Users

should be advised to consider the appropriate contextual information and be cautioned against misinterpretation.

Standard 15.13

Those who mandate testing programs should ensure that the individuals who interpret the test results to make decisions within the school or program context are qualified to assume this responsibility and proficient in the appropriate methods for interpreting test results.

Comment: When testing programs are used as a strategy for guiding interventions or instruction, professionals expected to make inferences leading to program improvement may need assistance in interpreting test results for this purpose.

The interpretation of some test scores is sufficiently complex to require that the user have relevant psychological training and experience. Examples of such tests include individually administered intelligence tests, personality inventories, projective techniques, and neuropsychological tests.

GLOSSARY

This glossary provides definitions of terms as used in this text. For many of the terms, multiple definitions can be found in the literature; also, technical usage may differ from common usage.

ability/trait parameter In item response theory (IRT), a theoretical value indicating the level of a test taker on the ability or trait measured by the test; analogous to the concept of true score in classical test theory.

ability testing The use of standardized tests to evaluate the current performance of a person in some defined domain of cognitive, psychomotor, or physical functioning.

absolute score interpretation The meaning of a test score for an individual or an average score for a defined group, indicating an individual's or group's level of performance in some defined criterion domain. By contrast, see *relative score interpretation*.

accommodation See *test modification*.

acculturation The process whereby individuals from one culture adopt the characteristics and values of another culture with which they have come in contact.

achievement levels/proficiency levels Descriptions of a test taker's competency in a particular area of knowledge or skill, usually defined as ordered categories on a continuum, often labeled from "basic" to "advanced," or "novice" to "expert," that constitute broad ranges for classifying performance. See *cut score*.

achievement testing A test to evaluate the extent of knowledge or skill attained by a test taker in a content domain in which the test taker had received instruction.

adaptive testing A sequential form of individual testing in which successive items, or sets of items, in the test are chosen based primarily on their psychometric properties and content, in relation to the test taker's responses to previous items.

adjusted validity/reliability coefficient A validity or reliability coefficient—most often, a product-moment correlation—that has been adjusted to offset the effects of differences in score variability, criterion variability, or the unreliability of test and/or criterion. See *restriction of range or variability*.

age equivalent The chronological age in a defined population for which a given score is the median (middle) score. Thus, if children 10 years and 6 months of age have a median score of 17 on a test, the score 17 is said to have an age equivalent of 10-6 for that population. See *grade equivalent*.

alternate forms Two or more versions of a test that are considered interchangeable, in that they measure the same constructs in the same ways, are intended for the same purposes, and are administered using the same directions. *Alternate forms* is a generic term used to refer to any of three categories. *Parallel forms* have equal raw score means, equal standard deviations, equal error structures, and equal correlations with other measures for any given population. *Equivalent forms* do not have the statistical similarity of parallel forms, but the dissimilarities in raw score statistics are compensated for in the conversions to derived scores or in form-specific norm tables. *Comparable forms* are highly similar in content, but the degree of statistical similarity has not been demonstrated. See *linkage*.

analytic scoring A method of scoring in which each critical dimension of performance

GLOSSARY

is judged and scored separately, and the resultant values are combined for an overall score. In some instances, scores on the separate dimensions may also be used in interpreting performance. See *holistic scoring*.

anchor test A common set of items administered with each of two or more different forms of a test for the purpose of equating the scores obtained on these forms.

assessment Any systematic method of obtaining information from tests and other sources, used to draw inferences about characteristics of people, objects, or programs.

attention assessment The process of collecting data and making an appraisal of a person's ability to focus on the relevant stimuli in a situation. The assessment may be directed at mechanisms involved in arousal, sustained attention, selective attention and vigilance, or limitation in the capacity to attend to incoming information.

automated narrative report See *computer-prepared test interpretation*.

back translation A translation of a test, which is itself a translation from an original test, back into the language of the original test. The degree to which a back translation matches the original test indicates the accuracy of the original translation.

battery A set of tests usually administered as a unit. The scores on the several tests usually are scaled so that they can readily be compared or used in combination for decision making.

bias In a statistical context, a systematic error in a test score. In discussing test fairness, bias may refer to construct underrepresentation or construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers.

See *predictive bias, construct underrepresentation, construct irrelevance*.

bilingual The characteristic of being relatively proficient in two languages.

calibration 1. In linking test score scales, the process of setting the test score scale, including mean, standard deviation, and possibly shape of score distribution, so that scores on a scale have the same relative meaning as scores on a related scale. **2.** In item response theory, the process of determining the parameters of the response function for an item.

certification A voluntary process, often national in scope, by which individuals who have been certified have demonstrated some level of knowledge and skill in an occupation. See *licensing, credentialing*.

classical test theory A psychometric theory based on the view that an individual's observed score on a test is the sum of a true score component for the test taker, plus an independent measurement error component.

classification accuracy The degree to which neither false positive nor false negative categorizations and diagnoses occur when a test is used to classify an individual or event. See *sensitivity and specificity*.

coaching Planned short-term instructional activities in which prospective test takers participate prior to the test administration for the primary purpose of improving their test scores. Coaching typically includes simple practice, instruction on test-taking strategies, and related activities. Activities that approximate the instruction provided by regular school curricula or training programs are not typically referred to as coaching.

coefficient alpha An internal consistency reliability coefficient based on the number

GLOSSARY

of parts into which the test is partitioned (e.g., items, subtests, or raters), the interrelationships of the parts, and the total test score variance. Also called *Cronbach's alpha* and, for dichotomous items, *KR 20*.

cognitive assessment The process of systematically gathering test scores and related data in order to make judgments about an individual's ability to perform various mental activities involved in the processing, acquisition, retention, conceptualization, and organization of sensory, perceptual, verbal, spatial, and psychomotor information.

composite score A score that combines several scores according to a specified formula.

computer-administered test A test administered by a computer. Questions appear on a computer-produced display, and the test taker answers by using a keyboard, "mouse" or other similar response device.

computer-based mastery test An adaptive test administered by computer that indicates whether or not the test taker has mastered a certain domain. The test is not designed to provide scores indicating degree of mastery, but only whether the test performance was above or below some specified level. Thus a *computer-based mastery test* is not simply a *mastery test* given by computer. See *mastery test*.

computer-based test See *computer-administered test*.

computer-generated test interpretation See *computer-prepared test interpretation*.

computer-prepared test interpretation A programmed, computer-prepared interpretation of an examinee's test results, based on empirical data and/or expert judgment.

computerized adaptive test An adaptive test administered by computer. See *adaptive testing*.

conditional measurement error variance The variance of measurement errors that affect the scores of examinees at a specified test score level; the square of the conditional standard error of measurement.

conditional standard error of measurement The standard deviation of measurement errors that affect the scores of examinees at a specified test score level.

confidence interval An interval between two values on a score scale within which, with specified probability, a score or parameter of interest lies. The term is also used in these standards to designate Bayesian credibility intervals that define the probability that the unknown parameter falls in the specified interval.

configural scoring rule A rule for scoring a set of two or more elements (such as items or subtests) in which the score depends on a particular pattern of responses to the elements.

construct The concept or the characteristic that a test is designed to measure.

construct domain The set of interrelated attributes (e.g., behaviors, attitudes, values) that are included under a construct's label. A test typically samples from this construct domain.

construct equivalence 1. The extent to which the construct measured by one test is essentially the same as the construct measured by another test. 2. The degree to which a construct measured by a test in one cultural or linguistic group is comparable to the construct measured by the same test in a different cultural or linguistic group.

construct irrelevance The extent to which test scores are influenced by factors that are irrelevant to the construct that the test is

GLOSSARY

intended to measure. Such extraneous factors distort the meaning of test scores from what is implied in the proposed interpretation.

construct underrepresentation The extent to which a test fails to capture important aspects of the construct that the test is intended to measure. In this situation, the meaning of test scores is narrower than the proposed interpretation implies.

construct validity A term used to indicate that the test scores are to be interpreted as indicating the test taker's standing on the psychological construct measured by the test. A construct is a theoretical variable inferred from multiple types of evidence, which might include the interrelations of the test scores with other variables, internal test structure, observations of response processes, as well as the content of the test. In the current standards, all test scores are viewed as measures of some construct, so the phrase is redundant with validity. The validity argument establishes the construct validity of a test. See *construct, validity argument*.

constructed response item An exercise for which examinees must create their own responses or products rather than choose a response from an enumerated set. Short-answer items require a few words or a number as an answer, whereas extended-response items require at least a few sentences.

content domain The set of behaviors, knowledge, skills, abilities, attitudes or other characteristics to be measured by a test, represented in a detailed specification, and often organized into categories by which items are classified.

content standard A statement of a broad goal describing expectations for students in a subject matter at a particular grade or at the completion of a level of schooling.

content validity A term used in the 1974 *Standards* to refer to a *kind* or *aspect* of validity that was "required when the test user wishes to estimate how an individual performs in the universe of situations the test is intended to represent" (p. 28). In the 1985 *Standards*, the term was changed to *content-related evidence* emphasizing that it referred to one type of evidence within a unitary conception of validity. In the current *Standards*, this type of evidence is characterized as "evidence based on test content."

convergent evidence Evidence based on the relationship between test scores and other measures of the same construct.

credentialing Granting to a person, by some authority, a credential, such as a certificate, license, or diploma, that signifies an acceptable level of performance in some domain of knowledge or activity.

criterion domain The construct domain of a variable used as a criterion. See *construct domain*.

criterion-referenced score interpretation
See *criterion-referenced test*.

criterion-referenced test A test that allows its users to make score interpretations in relation to a functional performance level, as distinguished from those interpretations that are made in relation to the performance of others. Examples of criterion-referenced interpretations include comparison to cut scores, interpretations based on expectancy tables, and domain-referenced score interpretations.

cross-validation A procedure in which a scoring system or set of weights for predicting performance, derived from one sample, is applied to a second sample in order to investigate the stability of prediction of the scoring system or weights.

GLOSSARY

cut score A specified point on a score scale, such that scores at or above that point are interpreted or acted upon differently from scores below that point. See *performance standard*.

derived score A score to which raw scores are converted by numerical transformation (e.g., conversion of raw scores to percentile ranks or standard scores).

diagnostic and intervention decisions Decisions based upon inferences derived from psychological test scores as part of an assessment of an individual that lead to placing the individual in one or more categories. See also *intervention planning*.

differential item functioning A statistical property of a test item in which different groups of test takers who have the same total test score have different average item scores or, in some cases, different rates of choosing various item options. Also known as DIF.

discriminant evidence Evidence based on the relationship between test scores and measures of different constructs.

documentation The body of literature (e.g., test manuals, manual supplements, research reports, publications, user's guides, etc.) made available by publishers and test authors to support test use.

domain sampling The process of selecting test items to represent a specified universe of performance.

empirical evidence Evidence based on some form of data, as opposed to that based on logic or theory. As used here, the term does not specify the type of evidence; this is in contrast to some settings where the term is equated with criterion-related evidence of validity.

equated forms Two or more test forms constructed to cover the same explicit content, to conform to the same statistical specifications, and to be administered under identical procedures (*alternate forms*); through statistical adjustments, the scores on the alternate forms share a common scale.

equating Putting two or more essentially parallel tests on a common scale. See *alternate forms*.

equivalent forms See *alternate forms*.

error of measurement The difference between an observed score and the corresponding true score or proficiency. See *standard error of measurement* and *true score*.

factor 1. Any variable, real or hypothetical, that is an aspect of a concept or construct. 2. In measurement theory, a statistical dimension defined by a factor analysis. See *factor analysis*.

factor analysis Any of several statistical methods of describing the interrelationships of a set of variables by statistically deriving new variables, called factors, that are fewer in number than the original set of variables.

factorial structure 1. The set of factors obtained in a factor analysis. 2. Technically, the correlation of each factor with each of the original variables from which the factors are derived.

fairness In testing, the principle that every test taker should be assessed in an equitable way. See chapter 7.

false negative In classification, diagnosis, or selection, an error in which an individual is assessed or predicted not to meet the criteria for inclusion in a particular group but in truth does (or would) meet these criteria. See *sensitivity* and *specificity*.

GLOSSARY

false positive In classification, diagnosis, or selection, an error in which an individual is assessed or predicted to meet the criteria for inclusion in a particular group but in truth does not (or would not) meet these criteria. See *sensitivity* and *specificity*.

field test A test administration used to check the adequacy of testing procedures, generally including test administration, test responding, test scoring, and test reporting. A field test is generally more extensive than a pilot test. See *pilot test*.

flag An indicator attached to a test score, a test item, or other entity to indicate a special status. A flagged test score generally signifies a score obtained in a modified, nonstandard test administration. A flagged test item generally signifies an item with undesirable characteristics, such as excessive differential item functioning.

functional equivalence In evaluating test translations, the degree to which similar activities or behaviors have the same functions in different cultural or linguistic groups.

gain score In testing, the difference between two scores obtained by a test taker on the same test or two equated tests taken on different occasions, often before and after some treatment.

generalizability coefficient A reliability index encompassing one or more independent sources of error. It is formed as the ratio of (a) the sum of variances that are considered components of test score variance in the setting under study to (b) the foregoing sum plus the weighted sum of variances attributable to various error sources in this setting. Such indices, which arise from the application of generalizability theory, are typically interpreted in the same manner as reliability coefficients. See *generalizability theory*.

generalizability theory An extension of classical reliability theory and methodology in which the magnitudes of errors from specified sources are estimated through the use of one or another experimental design, and the application of the statistical techniques of the analysis of variance. The analysis indicates the generalizability of scores beyond the specific sample of items, persons, and observational conditions that were studied.

grade equivalent The school grade level for a given population for which a given score is the median score in that population. See *age equivalent*.

high-stakes test A test used to provide results that have important, direct consequences for examinees, programs, or institutions involved in the testing.

holistic scoring A method of obtaining a score on a test, or a test item, based on a judgment of overall performance using specified criteria. See *analytic scoring*.

informed consent The agreement of a person, or that person's legal representative, for some procedure to be performed on or by the individual, such as taking a test or completing a questionnaire. The agreement, which is usually written, is made after the nature, possible effects, and use of the procedure has been explained.

intelligence test A psychological or educational test designed to measure an individual's level of cognitive functioning in accord with some recognized theory of intelligence.

internal consistency coefficient An index of the reliability of test scores derived from the statistical interrelationships of responses among item responses or scores on separate parts of a test.

GLOSSARY

internal structure In test analysis, the factorial structure of item responses or subscales of a test. See *factorial structure*.

inter-rater agreement The consistency with which two or more judges rate the work or performance of test takers; sometimes referred to as *inter-rater reliability*.

intervention planning The activity of a practitioner that involves the development of a treatment protocol.

inventory A questionnaire or checklist, usually in the form of a self-report, that elicits information about an individual's personal opinions, interests, attitudes, preferences, personality characteristics, motivations, and typical reactions to situations and problems.

item A statement, question, exercise, or task on a test for which the test taker is to select or construct a response, or perform a task. See *item prompt*.

item characteristic curve A mathematical function relating the probability of a certain item response, usually a correct response, to the level of the attribute measured by the item. Also called *item response curve*, or *item response function*, or *icc*.

item pool The aggregate of items from which a test or test scale's items are selected during test development, or the total set of items from which a particular test is selected for a test taker during adaptive testing.

item prompt The question, stimulus, or instructions that direct the efforts of examinees in formulating their responses to a constructed-response exercise.

item response theory (IRT) A mathematical model of the relationship between performance on a test item and the test taker's level of

performance on a scale of the ability, trait, or proficiency being measured, usually denoted as θ . In the case of items scored 0 / 1 (incorrect/correct response) the model describes the relationship between θ and the item mean score (P) for test takers at level θ , over the range of permissible values of θ . In most applications, the mathematical function relating P to θ is assumed to be a logistic function that closely resembles the cumulative normal distribution.

job analysis A general term referring to the investigation of positions or job classes to obtain descriptive information about job duties and tasks, responsibilities, necessary worker characteristics (e.g. knowledge, skills, and abilities), working conditions, and/or other aspects of the work.

job performance measurement The measurement of an incumbent's performance of a job. This may include a job sample test, an assessment of job knowledge, and possibly ratings of the incumbent's actual performance on the job.

job sample test A test of the ability of an individual to perform the tasks of which the job is comprised.

licensing The granting, usually by a government agency, of an authorization or legal permission to practice an occupation or profession. See also *certification*, *credentialing*.

linkage The result of placing two or more tests on the same scale, so that scores can be used interchangeably. Several linking methods are used: See *equating*, *calibration*, *moderation*, and *projection*, and *alternate forms*.

literature In this document, a term denoting accessible reports of research, such as books, articles published in professional journals, technical reports, and accessible versions of papers presented at professional meetings.

GLOSSARY

local evidence Evidence (usually related to reliability or validity) collected for a specific set of test takers in a single institution or at a specific location.

local norms Norms by which test scores are referred to a specific, limited *reference population* of particular interest to the test user (e.g., locale, organization, or institution); local norms are not intended as representative of populations beyond that setting.

local setting The organization or institution where a test is used.

low-stakes test A test used to provide results that have only minor or indirect consequences for examinees, programs, or institutions involved in the testing.

mandated tests Tests that are administered because of a mandate from an external authority.

mastery test 1. A criterion-referenced test designed to indicate the extent to which the test taker has mastered some domain of knowledge or skill. *Mastery is generally indicated by attaining a passing score or cut score.* 2. In some technical use, a test designed to indicate whether a test taker has or has not attained a prescribed level of mastery of a domain. See *cut score, computer-based mastery test.*

matrix sampling A measurement format in which a large set of test items is organized into a number of relatively short item sets, each of which is randomly assigned to a subsample of test takers, thereby avoiding the need to administer all items to all examinees in a program evaluation.

meta-analysis A statistical method of research in which the results from several independent, comparable studies are combined to determine the size of an overall effect or the degree of relationship between two variables.

moderation In test linking, the term moderation, used without a modifier, usually signifies statistical moderation, which is the adjustment of the score scale of one test, usually by setting the mean and standard deviation of one set of test scores to be equal to the mean and standard deviation of another distribution of test scores.

moderator variable In regression analysis, a variable that serves to explain, at least in part, the correlation of two other variables.

modification See *test modification.*

neuropsychodiagnosis Classification or description of inferred central nervous system status on the basis of neuropsychological assessment.

neuropsychological assessment A specialized type of psychological assessment of normal or pathological processes affecting the central nervous system and the resulting psychological and behavioral functions or dysfunctions.

norm-referenced test interpretation A score interpretation based on a comparison of a test taker's performance to the performance of other people in a specified *reference population*. See *criterion-referenced test.*

normalized standard score A derived test score in which a numerical transformation has been chosen so that the score distribution closely approximates a normal distribution, for some specific population.

norms Statistics or tabular data that summarize the distribution of test performance for one or more specified groups, such as test takers of various ages or grades. Norms are usually designed to represent some larger population, such as test takers throughout the country. The group of examinees represented by the norms is referred to as the *reference population*.

GLOSSARY

operational use The actual use of a test, after initial test development has been completed, to inform an interpretation, decision, or action based, in part, upon test scores.

outcome evaluation An evaluation of the efficacy of an intervention.

parallel forms See *alternate forms*.

percentile The score on a test below which a given percentage of scores fall.

percentile rank Most commonly, the percentage of scores in a specified distribution that fall below the point at which a given score lies. Sometimes the percentage is defined to include scores that fall at the point; sometimes the percentage is defined to include half of the scores at the point.

performance assessments Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied.

performance standard 1. An objective definition of a certain level of performance in some domain in terms of a cut score or a range of scores on the score scale of a test measuring proficiency in that domain. 2. A statement or description of a set of operational tasks exemplifying a level of performance associated with a more general content standard; the statement may be used to guide judgments about the location of a cut score on a score scale. The term often implies a desired level of performance. See *cut score*.

personality inventory An inventory that measures one or more characteristics that are regarded generally as psychological attributes or interpersonal proclivities or skills.

pilot test A test administered to a sample of test takers to try out some aspects of the test or test items, such as instructions, time limits, item response formats, or item response options. See *field test*.

policy The principles, plan, or procedures established by an agency, institution, organization, or government, generally with the intent of reaching a long-term goal.

portfolio In assessment, a systematic collection of educational or work products that have been compiled or accumulated over time, according to a specific set of principles.

precision of measurement A general term that refers to a measure's sensitivity to measurement error. See *standard error of measurement*, *error of measurement*.

practice analysis A general term referring to the investigation of a certain work position, or profession, to obtain descriptive information about the activities and responsibilities of the position and about the knowledge, skills, and abilities needed to engage in the work of the position. The concept is essentially the same as a job analysis but is generally preferred for professional occupations involving a great deal of individual decision making. See *job analysis*.

predictive bias The systematic under- or over-prediction of criterion performance for people belonging to groups differentiated by characteristics not relevant to criterion performance.

predictive validity A term used in the 1974 *Standards* to refer to a type of "criterion-related validity" that applies "when one wishes to infer from a test score an individual's most probable standing on some other variable called a criterion" (p. 26). In the 1985 *Standards*, the term *criterion-related validity* was changed to *criterion-related evidence*, emphasizing that it referred

GLOSSARY

to one type of evidence within a unitary conception of validity. The current document refers to "evidence based on relations to other variables" that include "test-criterion relationships." Predictive evidence indicates how accurately test data can predict criterion scores that are obtained at a later time.

program evaluation The collection and synthesis of systematic evidence about the use, operation, and effects of some planned set of procedures.

program norms See *user norms*.

projection In test scaling, a method of linking in which scores on one test (X) are used to predict scores on another test (Y). The projected Y score is the average Y score for all persons with a given X score. Like regression, the projection of test Y onto test X is different from the projection of test X onto test Y. See *linkage*.

proposed interpretation A summary, or a set of illustrations, of the intended meaning of test scores, based on the construct(s) or concept(s) the test is designed to measure.

protocol A record of events. A test protocol will usually consist of the test record and test scores.

psychodiagnosis Formalization or classification of functional mental health status based on psychological assessment. See *neuropsychodiagnosis*.

psychological assessment A comprehensive examination of psychological functioning that involves collecting, evaluating, and integrating test results and collateral information, and reporting information about an individual. Various methods may be used to acquire information during a psychological assessment: administering, scoring and interpreting tests and inventories; behavioral observation; client and third-party interviews; analysis of prior educational, occupational, medical, and psychological records.

psychological testing Any procedure that involves the use of tests or inventories to assess particular psychological characteristics of an individual.

random error An unsystematic error; a quantity (often observed indirectly) that appears to have no relationship to any other variable.

random sample See *sample*.

raw score The unadjusted score on a test, often determined by counting the number of correct answers, but more generally a sum or other combination of item scores. In item response theory, the estimate of test taker proficiency, usually symbolized $\hat{\theta}$, is analogous to a raw score although, unlike a raw score, its scaling is not arbitrary.

reference population The population of test takers represented by test norms. The sample on which the test norms are based must permit accurate estimation of the test score distribution for the reference population. The reference population may be defined in terms of examinee age, grade, or clinical status at time of testing, or other characteristics.

relative score interpretation The meaning of the test score for an individual, or the average score for a definable group, derived from the rank of the score or average within one or more reference distributions of scores. See *absolute score interpretation*.

reliability The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable, and repeatable for an individual test taker; the degree to which scores are free of errors of measurement for a given group. See *generalizability theory*.

GLOSSARY

reliability coefficient A unit-free indicator that reflects the degree to which scores are free of measurement error. The indicator resembles (or is) a product-moment correlation. In classical test theory, the term represents the ratio of true score variance to observed score variance for a particular examinee population. The conditions under which the coefficient is estimated may involve variation in test forms, measurement occasions, raters, scorers, or clinicians, and may entail multiple examinee products or performances. These and other variations in conditions give rise to qualifying adjectives, such as alternate-form reliability, internal consistency reliability, test-retest reliability, etc. See *generalizability theory*.

response bias A test taker's tendency to respond in a particular way or style to items on a test (i.e., acquiescence, social desirability, the tendency to choose 'true' on a true-false test) that yields systematic, construct-irrelevant error in test scores.

response process A component, usually hypothetical, of a cognitive account of some behavior, such as making an item response.

response protocol A record of the responses given by a test taker to a particular test.

restriction of range or variability Reduction in the observed score variance of an examinee sample, compared to the variance of the entire examinee population, as a consequence of constraints on the process of sampling examinees. See *adjusted validity/reliability coefficient*.

rubric See *scoring rubric*.

sample A selection of a specified number of entities called sampling units (test takers, items, etc.) from a larger specified set of possible

entities, called the population. A random sample is a selection according to a random process, with the selection of each entity in no way dependent on the selection of other entities. A stratified random sample is a set of random samples, each of a specified size, from several different sets, which are viewed as strata of the population.

scale **1.** The system of numbers, and their units, by which a value is reported on some dimension of measurement. Length can be reported in the English system of feet and inches or in the metric system of meters and centimeters. **2.** In resting, *scale* sometimes refers to the set of items or subtests used in the measurement and is distinguished from a test in the type of characteristic being measured. One speaks of a test of verbal ability, but a scale of extroversion-introversion.

scale score See *derived score*.

scaling The process of creating a scale or a scale score. Scaling may enhance test score interpretation by placing scores from different tests or test forms onto a common scale or by producing scale scores designed to support criterion-referenced or norm-referenced score interpretations. See *scale*.

score Any specific number resulting from the assessment of an individual; a generic term applied for convenience to such diverse measures as test scores, estimates of latent variables, production counts, absence records, course grades, ratings, and so forth.

scoring formula The formula by which the raw score on a test is obtained. The simplest scoring formula is "raw score equals number correct." Other formulas differentially weight item responses. For example, in an attempt to correct for guessing or nonresponse, zero weights may be assigned to nonresponses and negative weights to incorrect responses.

GLOSSARY

scoring rubric The established criteria, including rules, principles, and illustrations, used in scoring responses to individual items and clusters of items. The term usually refers to the scoring procedures for assessment tasks that do not provide enumerated responses from which test takers make a choice. Scoring rubrics vary in the degree of judgment entailed, in the number of distinct score levels defined, in the latitude given scorers for assigning intermediate or fractional score values, and in other ways.

screening test A test that is used to make broad categorizations of examinees as a first step in selection decisions or diagnostic processes.

security (of a test) See *test security*.

selection A purpose for testing that results in the acceptance or rejection of applicants for a particular educational or employment opportunity.

sensitivity In classification of disorders, the proportion of cases in which a disorder is detected when it is in fact present.

Spearman-Brown formula A formula derived within classical test theory that projects the reliability of a shortened or lengthened test from the reliability of a test of specified length.

specificity In classification of disorders, the proportion of cases for which a diagnosis of disorder is rejected when rejection is warranted.

speededness A test characteristic, dictated by the test's time limits, that results in a test taker's score being dependent on the rate at which work is performed as well as the correctness of the responses. The term is not used to describe tests of speed. Speededness is often an undesirable characteristic.

split-halves reliability coefficient An internal consistency coefficient obtained by using half the items on the test to yield one score and the other half of the items to yield a second, independent score. The correlation between the scores on these two half-tests, adjusted via the Spearman-Brown formula, provides an estimate of the alternate-form reliability of the total test.

stability The extent to which scores on a test are essentially invariant over time. Stability is an aspect of reliability and is assessed by correlating the test scores of a group of individuals with scores on the same test, or an equated test, taken by the same group at a later time.

standard error of measurement The standard deviation of an individual's observed scores from repeated administrations of a test (or parallel forms of a test) under identical conditions. Because such data cannot generally be collected, the standard error of measurement is usually estimated from group data. See *error of measurement*.

standard score A type of derived score such that the distribution of these scores for a specified population has convenient, known values for the mean and standard deviation. The term is sometimes used to signify a mean of 0.0 and a standard deviation of 1.0. See *derived score*.

standardization 1. In test administration, maintaining a constant testing environment and conducting the test according to detailed rules and specifications, so that testing conditions are the same for all test takers. 2. In test development, establishing scoring norms based on the test performance of a representative sample of individuals with which the test is intended to be used. 3. In statistical analysis, transforming a variable so that its standard deviation is 1.0 for some specified population or sample. See *standard score*.

GLOSSARY

- standards-based assessment** Assessments intended to represent systematically described content and performance standards.
- stratified coefficient alpha** A modification of coefficient alpha that renders it appropriate for a multi-factor test by defining the total score as the composite of scores on single-factor part-tests.
- stratified sample** See *sample*.
- systematic error** A consistent score component (often observed indirectly), not related to the test performance. See *bias*.
- technical manual** A publication prepared by test authors and publishers to provide technical and psychometric information on a test.
- test** An evaluative device or procedure in which a sample of an examinee's behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process.
- test developer** The person(s) or agency responsible for the construction of a test and for the documentation regarding its technical quality for an intended purpose.
- test development** The process through which a test is planned, constructed, evaluated, and modified, including consideration of content, format, administration, scoring, item properties, scaling, and technical quality for its intended purpose.
- test documents** Publications such as test manuals, technical manuals, user's guides, specimen sets, and directions for test administrators and scorers that provide information for evaluating the appropriateness and technical adequacy of a test for its intended purpose.
- test information function** A mathematical function relating each level of an ability or latent trait, as defined under item response theory (IRT), to the reciprocal of the corresponding conditional measurement error variance.
- test manual** A publication prepared by test developers and publishers to provide information on test administration, scoring, and interpretation and to provide technical data on test characteristics. See *user's guide*.
- test modification** Changes made in the content, format, and/or administration procedure of a test in order to accommodate test takers who are unable to take the original test under standard test conditions.
- test security** Limiting access to the specific content of a test to those who need to know it for test development, test scoring, and test evaluation. In particular, test items on secure tests are not published; unauthorized copying is forbidden by any test taker or anyone otherwise associated with the test. A secure test is not for publication in any form, in any venue.
- test specifications** A detailed description for a test, often called a test blueprint, that specifies the number or proportion of items that assess each content and process/skill area; the format of items, responses, and scoring rubrics and procedures; and the desired psychometric properties of the items and test such as the distribution of item difficulty and discrimination indices.
- test user** The person(s) or agency responsible for the choice and administration of a test, for the interpretation of test scores produced in a given context, and for any decisions or actions that are based, in part, on test scores.
- test-retest reliability** A reliability coefficient obtained by administering the same test a second time to the same group after a time interval and correlating the two sets of scores.

GLOSSARY

timed tests A test administered to a test taker who is allotted a strictly prescribed amount of time to respond to the test.

top-down A method of selecting the best applicants according to some numerical scale of suitability. Often, "best" is taken to mean "highest scoring on some test."

translational equivalence The degree to which the translated version of a test is equivalent to the original test. Translational equivalence is typically examined in terms of the language used, the scores produced, and the constructs measured by the translated version and the original test. See *back translation*.

true score In classical test theory, the average of the scores that would be earned by an individual on an unlimited number of perfectly parallel forms of the same test. In item response theory, the error-free value of test taker proficiency, usually symbolized by θ .

unidimensional Having only one dimension, or only one latent variable.

user norms Descriptive statistics (including percentile ranks) for a sample of test takers that does not represent a well-defined reference population, for example, all persons tested during a certain period of time, or a set of self-selected test takers. Also called program norms. See *norms*.

user's guide A publication prepared by the test authors and publishers to provide information on a test's purpose, appropriate uses, proper administration, scoring procedures, normative data, interpretation of results, and case studies. See *test manual*.

validation The process through which the validity of the proposed interpretation of test scores is investigated.

validity The degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test.

validity argument An explicit scientific justification of the degree to which accumulated evidence and theory support the proposed interpretation(s) of test scores.

validity generalization Applying validity evidence obtained in one or more situations to other similar situations on the basis of simultaneous estimation, meta-analysis, or synthetic validation arguments.

variance components In testing, variances accruing from the separate constituent sources that are assumed to contribute to the overall variance of observed scores. Such variances, estimated by methods of the analysis of variance, often reflect situation, location, time, test form, rater, and related effects.

vocational assessment A specialized type of psychological assessment designed to generate hypotheses and inferences about interests, work needs and values, career development, vocational maturity, and indecision.

weighted scoring A method of scoring a test in which the number of points awarded for a correct (or diagnostically relevant) response is not the same for all items in the test. In some cases, the scoring formula awards more points for one response to an item than for another.

INDEX

Numbers in this index refer to specific standard(s).

- Acceptable performance on credentialing test, 14.17
 - Based on knowledge and skills only, 14.17
- Accommodation, see "Test modifications"
- Achievement in instructional domain, 13.3
- Actuarial basis for recommendations and decisions, 12.17
- Adaptive testing procedures, 2.16
- Adequacy of fit, 3.9
- Adequacy of item or test performance, 4.21
- Adjusted validity/reliability coefficient, 1.18
- Administration, 2.18, 3.6, 3.9, 3.20-3.21, 5.1-5.7, 6.7-6.8, 6.11, 6.15, 8.1-8.3, 9.3, 9.5, 9.11, 10.1, 10.5-10.6, 10.8, 11.1, 11.3, 11.5, 11.9, 11.13, 11.16, 11.19, 11.22, 12.5, 12.8, 12.10-12.12, 13.6, 13.10-13.12, 13.16, 13.18, 15.10
 - Accommodations for examinees with disabilities, 2.18, 10.1, 10.8, 11.16
 - Adequate training of administrator, 12.8, 13.10, 13.12
 - Advance information, 8.2, 12.10, 15.10
 - Alternate methods, 6.11, 13.6
 - Clarity of directions, 3.20
 - Computer-administered tests, 2.8, 8.3, 13.18
 - Computer-scored tests, 13.18
 - Conditions, 3.9, 5.4, 8.1, 12.12
 - Consent forms, 6.15
 - Disruptions, 5.2
 - Examinee's most proficient language, 9.3
 - Guessing, 3.20
 - How to make responses, 5.5
 - Interpreters, 9.11
 - Minimize possibility of breaches in test security, 5.6
 - Modifications of standard procedures, 2.18, 5.2-5.3, 9.5, 11.19, 12.12
 - Monitoring, 5.4-5.5
 - Opportunity to practice using equipment, 5.5
 - Paper-and-pencil administration, 2.8, 8.3
 - Permissible variation in conditions, 3.21
 - Practice materials, 3.20, 8.1, 13.11
 - Protect security of test materials, 5.7, 11.9, 12.11
 - Questions from test takers, 3.20
 - Self-scored tests, 6.8
 - Special qualifications, 11.3
 - Standard administration instructions, 3.20, 12.8, 12.12, 13.10
 - Standardized instructions to test takers, 5.5
 - Standardized procedures, 5.1-5.2
 - Test taking strategies, 11.13
 - Time limits, 3.20, 10.6
 - User qualifications, 6.7, 13.12
- Advance information, 8.2, 8.4, 11.5, 11.13, 12.10, 14.16, 15.10
- Confidentiality protection, 8.2
- Consequences of misconduct, 8.2
- Rules and procedures to determine overall outcome of credentialing tests, 14.16
- Scoring criteria, 8.2
- Test taking strategies, 8.2, 11.13
- Testing policy, 8.2, 12.10, 15.10
- Time limits, 8.2, 12.10
- To test takers, 8.2, 8.4, 12.10
- Use of test scores, 8.2, 12.10, 15.10
- Advancement, 9.8
- Alternate forms, see "Test forms"
- Anchor test, 4.11, 4.13
 - Psychometric characteristics, 4.13
 - Representativeness, 4.13
- Arbitration of disputes, 8.11
- Attenuation, correction for, 1.18, 2.6
- Attrition rates, 15.4
- Benchmarks**, 13.19
- Bias, 7.3-7.4, 7.12, 11.24, 12.2
- Calibration**, 4.15, 5.12, 12.12
- Case studies, 6.10, 10.12
- Categorical decisions, 2.15
- Census-type testing programs, 11.24
- Change scores, 13.17, 15.3
- Characteristics of job, 14.10, 14.12
- Cheating, 8.2, 8.7, 8.10-8.11, 11.11,
- Classification, 2.14, 3.7, 3.22, 4.9, 4.19, 14.7, 14.8
 - Employment, 14.7, 14.8
 - Of constructed responses, 3.22
 - Of examinees, 4.9, 4.19
- Classification consistency, 2.15
- Clinical and counseling settings, 11.20
- Coaching, 1.9
- Coding, 3.22
- Collateral information, 12.18
- Combining tests, 12.4-12.5
 - Addressing complex diagnoses, 12.5
 - Justification for interpretation, 12.4
 - Rationale, 12.4
- Comparability, 4.10, 7.8, 9.4, 9.9, 10.4, 10.11, 13.8, 14.11
 - Across groups, 7.8
 - Job content factors, 14.11
 - Modifications for individuals with disabilities, 10.4
 - Multiple-language versions of test, 9.9
 - Score, 4.10, 9.4, 10.11, 13.8
- Computer-administered tests, 2.8, 5.5, 6.11, 8.2-8.3, 13.18
 - Documentation of design, 13.18
 - Documentation of scoring algorithms, 13.18
 - Methods for scoring and classifying, 13.18

INDEX

- Computer-based testing, 13.18
 - Construct-irrelevant variance, 13.18
- Computer-generated interpretations, 5.11, 6.12, 11.21, 12.15
 - Cut scores, 6.12
 - Empirical basis, 5.11
 - Limitations, 5.11, 11.21, 12.15
 - Norms, 12.15
 - Quality, 12.15
 - Rationale, 5.11
 - Sources, 5.11
- Computerized adaptive tests, 3.12, 4.10, 8.3
 - Documentation, 3.12
 - Rationale, 3.12, 4.10
 - Supporting evidence, 3.12
- Concordance tables, 4.14
- Conditional standard errors of measurement, 2.14
- Confidence interval, 2.2
- Confidentiality protection, 8.2, 8.6, 12.11
- Conflict of interest, 12.2
- Consequences of misconduct, 8.2
- Consequences of test use, 1.24
- Consistency of scores, 2.4
- Construct description, 1.2
- Construct equivalent tests, 7.2, 13.6
- Construct-irrelevant variance, 7.2, 7.10, 12.19, 13.18
- Construct overlap, 13.8
- Construct representation, 7.11
- Construct underrepresentation, 7.10
- Content domain, 1.6, 3.11, 7.3, 13.5, 14.8, 14.10, 14.14
 - Job, 14.10
- Content specifications, 1.6
- Context effects, 2.17, 4.15, 13.15
- Controlling item exposure, 3.12
- Convergent evidence, 12.18
- Converted scores, 4.16
 - Possible nonequivalence in revisions, 4.16
- Copyright, 8.7, 11.8-11.9, 12.11
 - Infringement, 8.7
 - Protection, 11.8-11.9, 12.11
- Copyright date, 6.14
- Credentialing testing, 9.8, 14.14-14.17
 - Credential-worthy performance in an occupation, 14.14
 - Level of performance required for passing, 14.17
 - Licensure and certification, 14.15
- Criterion construct domain, 14.12
- Criterion-referenced interpretation, 4.1, 4.9
 - Empirical basis, 4.9
 - Rationale, 4.9
- Criterion-referenced testing programs, 3.4, 14.2
- Cross-validation studies, 3.10
- Cultural differences, 9.1-9.11
- Curriculum standards, 13.3
- Cut scores, 2.14-2.15, 4.4, 4.11, 4.19-4.21, 6.5, 6.12, 13.6, 14.17
 - Expert judgment, 4.21
 - Legal requirements, 4.19
 - Pass/fail, 4.21
 - Procedures for establishing, 4.19
 - Proficiency categories, 4.21
 - Rationale, 4.19
 - Relation of test performance to relevant criteria, 4.20
- Decision making, 11.4, 12.17, 13.5, 13.7-13.9, 13.13, 14.7, 14.13, 14.15-14.16
 - Actuarial basis, 12.17
 - Certification, 14.15
 - Classification, 11.4, 13.7
 - Construct overlap, 13.8
 - Desired student outcomes, 13.9
 - Diagnosis, 11.4
 - Educational placement, 13.9
 - Graduation, 13.5
 - Integrating information from multiple tests and sources, 14.13
 - Job classifications, 14.7
 - Pass/fail, 14.16
 - Promotion, 13.5, 13.9
 - School context, 13.13
 - Selection, 11.4
 - Validity, 11.4, 13.7
- Defined domain, 3.11
- Derived score scales, 4.1
 - Intended interpretation, 4.1
 - Limitations, 4.1
 - Meanings, 4.1
- Derived scores, 2.2, 3.22, 4.2, 4.7, 6.5
- Descriptive statistics, 2.4
- Difference scores, 13.8
 - Standardized tests, 13.8
- Differential diagnosis, 12.6
 - Ability to distinguish between multiple groups of concern, 12.6
- Differential item functioning (DIF), 7.3
- Differential prediction hypothesis, 7.6
- Disabilities (*resting individuals with*), see "Testing individuals with disabilities"
- Diversity, 6.10, 9.1-9.8, 9.10-9.11, 10.1-10.12, 11.22-11.23
 - Individuals with disabilities, 10.1-10.12, 11.23
 - Linguistic, 9.1-9.8, 9.10-9.11, 11.22-11.23
- Documentation, see "Publisher materials/responsibilities"
- Educational testing programs, 8.10-8.13, 9.3, 11.20, 13.1-13.19, 15.7, 15.12-15.13
 - Average of summary scores for groups, 13.19, 15.12
 - Educational placement, 13.9
 - Graduation, 13.5-13.6
 - Group differences in test scores, 13.15
 - Guiding instructions, 13.13, 15.13

INDEX

- Mandated tests, 15.7, 15.13
- Promotion, 13.5-13.6, 13.9
- Qualifications of administrators, 13.10
- Qualifications of scorers, 13.10
- Score reports, 13.14
- Special needs identification, 13.7
- Standards for mastery, 13.5-13.6
- Validity of score inferences as time passes, 13.16
- Effects of disabilities on test performance, 10.2
- Empirical evidence, 4.20, 7.6, 9.7, 10.5, 12.16, 13.9, 14.4-14.5, 15.8
 - Contaminants and artifacts, 14.5
 - Supporting basis for expecting specific outcomes, 15.8
- Employment testing, 9.8, 14.1-14.13
 - Classification, 14.8
 - Job analysis, 14.4, 14.6
 - Job classification decisions, 14.7
 - Objectives, 14.1
 - Personnel selection, 14.12
 - Prediction, 14.1, 14.4
 - Predictor-criterion relationships, 14.2-14.6
 - Promotion, 14.8-14.9
 - Screening, 14.1
 - Selection, 14.8-14.9
- Equated forms, 4.11
- Equating procedures, 4.11
- Equating studies, 4.11-4.13
 - Anchor test design, 4.13
 - Characteristics of anchor tests or linking items, 4.11
 - Classical, 4.13
 - Design, 4.11
 - Examinee samples, 4.11
 - IRT-based, 4.13
 - Statistical equivalence of examinee groups, 4.12
 - Statistical methods used, 4.11
- Error of measurement, 14.5
- Error variances, 2.5
- Ethics, 12.2, 12.10
- Evaluation, 15.2
 - Relevance of test to program goals, 15.2
- Examinee performance, 2.8-2.9
- Examinee subgroups, 7.1-7.4, 7.6, 7.10-7.12, 11.24
- Expert judgment, 1.7, 3.5-3.7, 3.11, 3.13, 4.19, 4.21, 14.9
 - Cut scores, 4.21
 - Demographic characteristics of judges, 3.5-3.6
 - Job task content, 14.9
 - Qualification of judges, 3.5-3.6
 - Relevant experiences of judges, 3.5-3.6
 - Standard setting, 4.19
- Expert review, 3.5
 - Process, 3.5
 - Purpose, 3.5
 - Results, 3.5
- Extended response items, 3.14
- Fairness, 7.1-7.12, 8.1, 8.11, 9.5, 10.11, 13.5-13.6
 - Absence of bias, 7.3-7.4, 7.12
 - Equality of testing outcomes for examinee subgroups, 7.8, 7.10-7.11
 - Equitable treatment of all examinees, 7.1-7.4, 7.8, 7.12, 8.1, 9.5, 10.11
 - Opportunity to learn, 7.10, 13.5-13.6
- Fatigue, 10.6
- Field tests, 3.8-3.9
- Flagged test score, 9.5, 10.11
- Forms, see "Test forms"
- Gain scores, 13.17, 15.3
 - Report of technical qualities, 13.17, 15.3
- Generalizability, 2.5, 2.10, 3.11, 12.16, 13.3
- Group-level information, 5.12, 11.24, 13.15, 15.12
 - Aggregating results, 5.12
 - Cautions against misrepresentations, 15.12
 - Differences, 13.15, 15.12
- Group means, 4.8
- Group performance measure, 2.20
- Group testing programs, 12.9
 - Professional supervisor responsibilities, 12.9
- Individual testing, 12.3, 12.18-12.19, 13.13
- Informed choice, 8.3
- Informed consent, 8.4-8.5
 - Exceptions, 8.4
- Integrity of test results, 15.9
- Inter-item correlation, 3.3
- Interpretation of individual item responses, 1.10
- Interpretation of test scores, see "Score interpretation"
- Interpreters, 9.11
 - Qualifications, 9.11
- Interpretive material for local release, 5.10, 15.13
 - Common misinterpretations, 5.10
 - How scores will be used, 5.10
 - Precision of scores, 5.10
 - Simple language, 5.10
 - What scores mean, 5.10
 - What test covers, 5.10
- Inter-rater agreement, 3.23
- Investigation of test taker misconduct, 8.10-8.12
- Irrelevant variance, 3.17
- Item development, 3.7
- Item evaluation, 3.9
 - Psychometric properties, 3.9
 - Sample description, 3.9
- Item pool, 4.17, 6.4
- Item response theory (IRT), 2.16, 3.9
 - Ability or trait parameter, 2.16
 - Item parameter estimates, 2.16, 3.9
- Item review, 3.7
- Item selection, 3.7, 3.9-3.10, 3.12
 - Empirical relationships, 3.10
 - Item difficulty, 3.9

INDEX

- Item discrimination, 3.9
- Item information, 3.9
- Procedures, 3.12
- Subsets of items, 3.12
- Tendency to select by chance, 3.10
- Item tryouts, 3.7-3.8
- Item weights, 3.13
 - Based on empirical data, 3.13
 - Based on expert judgment, 3.13
- Job analysis, 14.6, 14.8, 14.11, 14.14
- Job content domain, 14.10
 - Abilities, 14.10
 - Knowledge, 14.10
 - Skills, 14.10
 - Tasks, 14.10
- Labels, 8.8
 - Least stigmatizing, 8.8
- Language differences (testing individuals with), 9.1-9.11, 11.22
 - Appropriateness of tests, 9.1, 11.22
- Language proficiency, 9.3, 9.8, 9.10, 11.22
 - Bilingual, 9.3
 - Communicative abilities, 9.10
 - Examinees, 9.3, 9.10
 - Multiple languages, 9.3
 - Required level for occupations, 9.8
- Large-scale testing programs, 5.3, 5.6, 5.12
- Learning opportunity changes, 13.15
- Legally mandated testing, 8.4
- Licensure and certification, 8.7, 8.10-8.13, 9.8, 14.14-14.17
 - Knowledge and skills necessary, 14.14
 - Purpose of program, 14.14
- Limitations of test scores, 11.2
- Linguistic ability, 7.7, 11.23
- Linguistic characteristics of examinees, 9.1-9.3, 9.5-9.6, 11.22
- Linguistic subgroups, 9.2
- Linkage, 4.15, 14.12
- Local scorers, see "Scorers"
- Logical evidence, 9.7
- Mandated testing programs**, 13.1, 15.7, 15.13
 - Description of ways results will be used, 13.1, 15.7, 15.13
 - Negative consequences, 13.1, 15.7, 15.13
- Mastery of skills, 13.6
- Matrix sampling, 2.20, 5.12, 15.6
- Measurement error, 13.8, 13.14
- Meta-analysis, 1.20, 1.21
- Moderator variables, 7.6
- Modifications, see "Test modifications"
- Monitoring, 5.4-5.5, 5.9, 12.8-12.9
 - Administration, 5.4-5.5, 12.8
 - Scoring, 5.9, 12.8-12.9
- Motivation of test takers, 15.4
- Multidisciplinary evaluation, 10.12
- Multimedia testing, 13.18
 - Documentation of design, 13.18
 - Documentation of scoring algorithms, 13.18
 - Methods of scoring and classifying, 13.18
- Multiple-aptitude test batteries, 13.8
 - Comparing scores from test components, 13.8
- Multiple-language tests, 8.3
- Multiple-purpose tests, 13.2, 15.1
 - Appropriate technical evidence for each purpose, 13.2, 15.1
- Normative data, 6.4-6.5, 13.16
 - Norming population, 6.4
 - Years of data collection, 6.4, 13.16
- Norming studies, 4.6
 - Dares of testing, 4.6
 - Descriptive statistics, 4.6
 - Participation rates, 4.6
 - Population, 4.6
 - Sampling procedures, 4.6
 - Weighting of sample, 4.6
- Norm-referenced interpretation, 4.1, 4.9, 13.13, 13.16
- Norm-referenced testing programs, 3.4
- Norms, 2.12, 3.19, 4.2, 4.5-4.8, 4.15, 4.18, 10.9, 11.19, 12.3, 12.12, 12.18, 13.4, 13.8, 13.13
 - Group means, 4.8
 - Individuals with disabilities, 10.9
 - Local, 4.7, 13.4
 - Precision, 4.6
- Outcome monitoring**, 15.5, 15.8
 - Basis for expecting outcome, 15.8
- Outcome of credentialing tests, 14.16
- Pass/fail, 14.16-14.17
 - Level of performance required, 14.16-14.17
- Performance assessments, 3.14
- Pilot testing, 10.3
- Policy studies, 15.2, 15.4-15.5, 15.11-15.12
 - Release of test results, 15.11-15.12
 - Suitability of test, 15.2
- Policy makers, 7.9, 15.11
 - Educational, 7.9
 - Public, 7.9
 - Social, 7.9
- Populations, 1.2, 1.5, 3.6, 3.8, 4.5-4.7, 6.4, 7.1, 7.3, 11.1, 11.16, 11.24, 12.3, 12.8, 12.16, 13.4, 13.8, 13.15, 15.5-15.6
 - Background of test taker, 12.3
 - Census-type testing programs, 11.24
 - Characteristics of test taker, 12.3

INDEX

- Cultural differences, 13.15
- Descriptions, 2.20, 15.6
- Gradual changes in demographic characteristics, 11.16
- Representativeness, 1.5, 12.16, 13.4, 15.5
- Subgroup differences, 7.1, 7.3, 13.15
- Practice effects, 1.9
- Precision of scores, 2.4
- Prediction, 14.1, 14.4, 14.6-14.7
 - Absenteeism, 14.4
 - Job behavior, 14.1
 - Job-relevant training, 14.4
 - Job success, 14.7
 - Turnover, 14.4
 - Work behaviors, 14.4
 - Work output, 14.4
- Predictor construct domain, 14.12
- Predictor-criterion relationships, 14.2-14.6
 - Grounded in research, 14.2
- Pretest/posttest scores, 13.17, 15.3
 - Change scores, 13.17, 15.3
 - Gain scores, 13.17, 15.3
- Privacy protection, 11.14
- Procedural protections, 8.12-8.13
- Proctors, 11.11
- Professional competence, 12.1, 12.5, 12.8, 12.10-12.11, 13.12-13.13
 - Credentialing, 12.1
 - Educational, 12.1
 - Experience, 12.1
 - Supervised training, 12.1
- Program evaluation, 2.18, 2.20, 15.1-15.13
 - Eliminate practices designed to raise test scores, 15.9
 - Interpretation and release of results, 15.13
 - Suitability of test to program goals, 15.2
- Program goals, 15.2
- Program monitoring, 2.16
- Promotion, 14.8-14.9
 - Employment, 14.8-14.9
- Psychological testing, 12.1-12.20
 - Complex diagnoses, 12.5
 - Diagnosis, 12.6-12.7
 - Diagnostic sensitivity and specificity, 12.5
 - Individual testing, 12.3
 - Interpretive remarks, 12.13
 - Potential inferences described as hypotheses, 12.13
 - Using tests in combination, 12.4-12.5
- Publisher materials/responsibilities, 1.1-1.3, 2.11-2.12, 3.1-3.5, 3.9-3.13, 3.15, 3.19-3.27, 4.1-4.6, 4.11, 4.14-4.16, 4.18-4.19, 5.1, 5.10, 5.14, 6.1-6.15, 7.3-7.4, 7.9-7.10, 8.1-8.2, 9.4, 9.6-9.7, 10.4-10.5, 10.7-10.8, 11.1, 11.3-11.4, 11.7-11.9, 11.13, 12.4
 - Administration procedures, 5.1
 - Amending, revising, or withdrawing test, 3.25, 6.13
 - Applicability of test to non-native speakers, 9.6
 - Case studies, 6.10
 - Cautions against misuses, 6.3, 11.7, 11.8
 - Computer-generated interpretations, 6.12
 - Consent forms, 6.15
 - Copyright date, 6.14
 - Corrected score report, 5.14
 - Criteria for scoring, 3.20
 - Directions for administration, 3.19
 - Directions to test takers, 3.3, 8.1
 - Documentation of procedures used to modify test, 10.5
 - Documentation without compromising security, 3.12, 11.18
 - Expected level of scorer agreement and accuracy, 3.24
 - Foreign language translation or adaptation procedures, 6.4
 - General information, 6.15
 - Identification of related course or curriculum, 6.6
 - Information to policy makers, 7.9, 11.18
 - Instructions for using rating scales, 3.22
 - Instructions to test takers, 3.20
 - Interpretation of scores, 1.9, 1.12
 - Interpretive material, 5.10, 6.8, 6.10
 - Linguistic modifications, 9.4
 - Modified forms, 10.8
 - Norming studies, 4.6, 6.4
 - Norms, 4.2, 4.5
 - Practice or sample questions or tests, 3.20, 8.1
 - Procedures for test administration and scoring, 3.3
 - Qualifications to administer and score test, 6.7
 - Rationale, 11.4
 - Rationale for modifications, 10.4
 - Recommendations and cautions regarding modifications, 10.4
 - Reliability data, 2.11-2.12, 6.5
 - Renorming with sufficient frequency, 4.18
 - Research to avoid bias, 7.3
 - Revisions and implications on test score interpretation, 3.26, 6.13
 - Sample material, 3.20
 - Score reports, 1.10
 - Scoring criteria, 3.22
 - Scoring procedures, 5.1
 - Security, 11.8-11.9
 - Sensitivity reviews, 7.4
 - Statements regarding research-use-only tests, 3.27
 - Statistical descriptions and analyses
 - Suggestions to use tests in combination, 12.4
 - Summaries of cited studies, 6.9

INDEX

- Supplemental material, 6.1
- Technical documentation, 4.2, 4.6, 4.19
- Technical manual, 6.1, 10.5
- Test bulletin (advance information), 8.2
- Test directions, 3.15
- Test manual, 1.10, 3.1, 4.16, 6.1-6.2, 6.4, 9.4, 10.4-10.5, 11.3
- Test taking strategies, 11.13
- Training materials for scorers, 3.23-3.24
- Translation information, 9.7
- User's guides, 6.1
- Validity information, 6.5
- Purpose of test, 3.2, 3.6, 8.1, 11.1-11.2, 11.5, 11.16, 11.24, 13.2-13.3, 13.7, 13.12, 14.14
- Range restriction**, 14.5
- Rationale, 1.1, 6.3, 9.4
- Raw scores, 4.4, 6.5
 - Intended interpretations, 4.4
 - Limitations, 4.4
 - Meanings, 4.4
- Reading ability, 7.7
- Relationship between test scores, 13.8-13.9, 13.12
- Release of summary test results to public, 11.17-11.18, 15.11
 - Policy for timely release, 11.17
 - Provision of supplemental explanations, 11.18, 15.11
- Reliability, 2.1-2.20, 3.3, 3.19, 3.23, 5.12, 9.1, 9.7, 9.9, 11.1-11.2, 11.19, 12.13, 13.8, 13.12, 14.15, 15.6
 - Alternate-form reliability estimate, 2.9
 - Analyses for scores produced under major variations, 2.18
 - Data for major populations, 2.11
 - Data for separate grades and age groups, 2.12
 - Data for subpopulations, 2.11
 - Decision reliability, 14.15
 - Difference scores, 13.8
 - Error variance estimates, 2.10
 - Estimates, 2.1, 2.9
 - Generalizability coefficient, 2.5
 - Inter-rater consistency, 2.10
 - Language differences, 9.1
 - Local reliability data, 2.12
 - Long and short versions of a test, 2.17
 - Rate of work, 2.8-2.9
 - Reliability estimation procedures, 2.7
 - Reported for level of aggregation, 5.12
 - Sampling procedures, 15.6
 - Scorer, 3.23
 - Sources of measurement error, 2.10
 - Speededness, see "Rate of work"
 - Systematic variance, 2.8
 - Test comparability, 9.9
 - Test-retest reliability estimate, 2.9
 - Translations of a test, 9.7
 - Within-examinee consistency, 2.10
- Reliability coefficients, 2.5-2.6, 2.11-2.12
 - Alternate-form coefficients, 2.5
 - Internal consistency coefficients, 2.5
 - Restriction of range or variability adjustment, 2.6
 - Test-retest or stability coefficients, 2.5
- Replicability, 12.12
- Research use only tests, 3.27
- Response format, 2.8, 3.6, 3.14, 3.22, 4.21, 5.1, 5.5, 11.13, 12.12
 - Constructed, 2.8, 3.22, 4.21
 - Extended-response, 3.14
 - Unstructured, 12.12
- Restriction of range or variability, 1.18, 2.6
- Retention policy, 5.15-5.16, 8.6, 11.5, 15.10
 - Confidentiality, 8.6
 - Data transmission security, 8.6
 - Protection from improper disclosure, 8.6
 - Valid use of information, 5.16, 15.10
- Retest opportunity, 11.12, 12.10, 13.6
- Rights of test taker, 8.10-8.13, 11.10-11.12, 12.20, 13.6
 - Appeal and representation by counsel, 11.11
 - Retest opportunity, 11.12, 13.6
- Rubric, see "Scoring rubric"
- Sample representativeness**, 3.8
- Sampling procedures, 2.4, 3.8, 3.10, 14.6, 15.6
- Scale development procedures, 6.4
- Scale stability, 4.17
 - Over time, 4.17
- Scales, 4.2
- Scaling, 3.22
- Score comparability, 4.10, 9.4, 10.11, 13.4
- Score conversions, 4.14
 - Limitations, 4.14
- Score differences, 2.3
- Score equivalence, 4.10-4.11
 - Direct evidence, 4.10
 - Equating procedures, 4.11
 - Intended uses, 4.10
- Score integrity, 5.6
- Score interpretation, 1.1-1.2, 1.9, 1.12, 1.23, 2.11, 3.4, 3.14, 3.16, 3.18, 3.25-3.26, 4.1, 4.3-4.4, 4.6-4.7, 4.10, 4.16, 4.18-4.20, 5.1, 5.10-5.11, 5.14, 6.3, 6.5, 6.7-6.8, 6.10-6.12, 7.1-7.5, 7.8, 8.7, 8.9, 9.2, 9.5-9.7, 9.9, 10.4-10.5, 10.7, 10.9, 10.11, 11.1, 11.3, 11.5-11.6, 11.15, 11.17-11.18, 11.20, 11.22, 12.9, 12.13, 12.19, 13.3, 13.7-13.9, 13.12-13.15, 14.13, 14.16, 15.11-15.13
 - Absolute, 3.4
 - Affected by revisions, 3.26, 4.16
 - Alternate explanations for test taker's performance, 7.5, 11.20, 12.19, 13.7
 - Case studies, 6.10

INDEX

- Computer-generated interpretations, 5.11, 6.12
- Contextual information, 13.15, 15.12
- Cut scores, 4.19-4.20, 6.5
- Difference scores, 13.8
- Effects of modifications for individuals with disabilities, 10.7
- Flagged scores, 9.5, 10.11
- Inferences within subpopulations, 2.11, 7.3-7.4
- Interpretive material for local release, 5.10, 11.17-11.18, 13.12-13.14, 15.11
- Item level information, 6.5
- Linguistically diverse examinees, 9.2, 9.6, 11.22
- Material error requires corrected score report, 5.14
- Modifications for individuals with disabilities, 10.4
- Norms, 4.6, 10.9
- Potential misinterpretations, 11.15, 13.14-13.15, 15.12
- Relative, 3.4
- Score equivalence, 4.10
- Scores obtained under alternate conditions, 6.11
- Self-scored tests, 6.8
- Short form, 3.16
- Special qualifications, 11.3
- Speed component appropriateness, 3.18
- Subgroup differences, 7.1, 7.8
- Translated tests, 9.7
- Valid inferences for examinee subgroups, 7.2
- Validity jeopardized by departure from standard procedures, 5.1
- Weighted scoring, 14.16
- Score reporting, 2.17, 5.13-5.16, 6.12, 7.8, 8.4-8.6, 8.8-8.11, 8.13, 9.4-9.5, 11.6, 11.12, 11.14, 11.17-11.18, 12.9, 12.15, 12.19-12.30, 13.16-13.17, 13.19, 15.3, 15.10-15.11
 - Age of norms used for reporting, 13.16
 - Anonymity for researchers, 8.5
 - Cancellation or withdrawal of scores, 8.11
 - Categorical decisions, 8.8
 - Change scores, 13.17, 15.3
 - Computer-generated interpretations, 6.12, 12.15
 - Conditions for disclosure, 11.14
 - Confidentiality, 5.13, 8.4-8.5, 8.9
 - Corrected score report, 5.14
 - Date of test administration, 13.16
 - Delays because of possible irregularities, 8.10
 - Description and analysis of alternate hypotheses or explanations, 12.19
 - Exam retakes, 11.12
 - Flagged test scores, 9.5
 - Format appropriate for recipient, 11.6, 12.9, 12.20, 13.14, 13.19, 15.11
 - Gain scores, 13.17, 15.3
 - Invalidation of score, 8.13
 - Linguistically modified tests, 9.4
 - Public reporting for groups, 7.8, 11.17-11.18, 13.19, 15.11
 - Request for review or revision of scores, 8.13
 - Retention of individual data, 5.15, 8.6, 15.10
 - Waiver of access, 8.9
- Score scales, 4.1-4.4, 4.9
 - Age-equivalent scores, 4.1
 - Criterion-referenced interpretation, 4.1-4.2, 4.9
 - Derived scores, 4.1, 4.4, 4.9
 - Forewarning of potential specific misinterpretations, 4.3
 - Grade-equivalent scores, 4.1
 - Norm-referenced interpretation, 4.1-4.2, 4.9
 - Percentile ranks, 4.1
 - Raw scores, 4.1, 4.4, 4.9
 - Standard score scales, 4.1
- Scorers, 2.12, 3.22-3.24, 5.9, 6.7, 12.8, 13.10
 - Accuracy, 3.24, 13.10
 - Agreement, 3.24
 - Feedback, 5.9
 - Local, 2.12, 3.22, 3.24
 - Monitoring, 5.9
 - Qualifications, 3.23, 6.7, 13.10
 - Reliability, 3.23
 - Retraining or dismissing, 5.9
 - Scorer judgment, 3.24, 5.9
 - Selecting, 3.23
 - Training, 3.23, 12.8, 13.10
- Scores, types
 - Composite scores, 1.12, 2.1, 2.7, 14.16
 - Subscores, 1.12, 2.1
- Scoring criteria, 3.14, 5.9, 8.2, 12.11
- Scoring errors, 5.8, 11.10
- Scoring procedures, 3.14, 5.1-5.2, 5.8-5.9
- Scoring rubrics, 3.23-3.24, 5.9
- Scoring services, 5.8, 6.12
- Screening, 11.5, 13.7, 14.1
 - Screening in, 14.1
 - Screening out, 14.1
- Selection, 2.14, 9.8, 14.8-14.9, 14.11-14.12
 - Employee, 14.8-14.9, 14.11-14.12
- Selection tests, 13.8
 - Comparing scores, 13.8
- Self-scored tests, 6.8
- Standard error of the difference score, 13.8, 13.17, 15.3
- Standard error of the group mean, 2.19
 - Variability due to measurement error, 2.19
 - Variability due to sampling, 2.19
- Standard errors of ability scores, 2.16
- Standard errors of equating functions, 4.11
- Standard errors of measurement, 2.1-2.3, 2.5, 2.11-2.12, 2.14, 6.5, 13.8, 14.15
 - Conditional, 2.2
 - Overall, 2.2
 - Repeated-measurements approach, 2.15
- Standard setting, 4.19-4.20
- Standardization, 3.20
- Standards for mastery, 13.5

INDEX

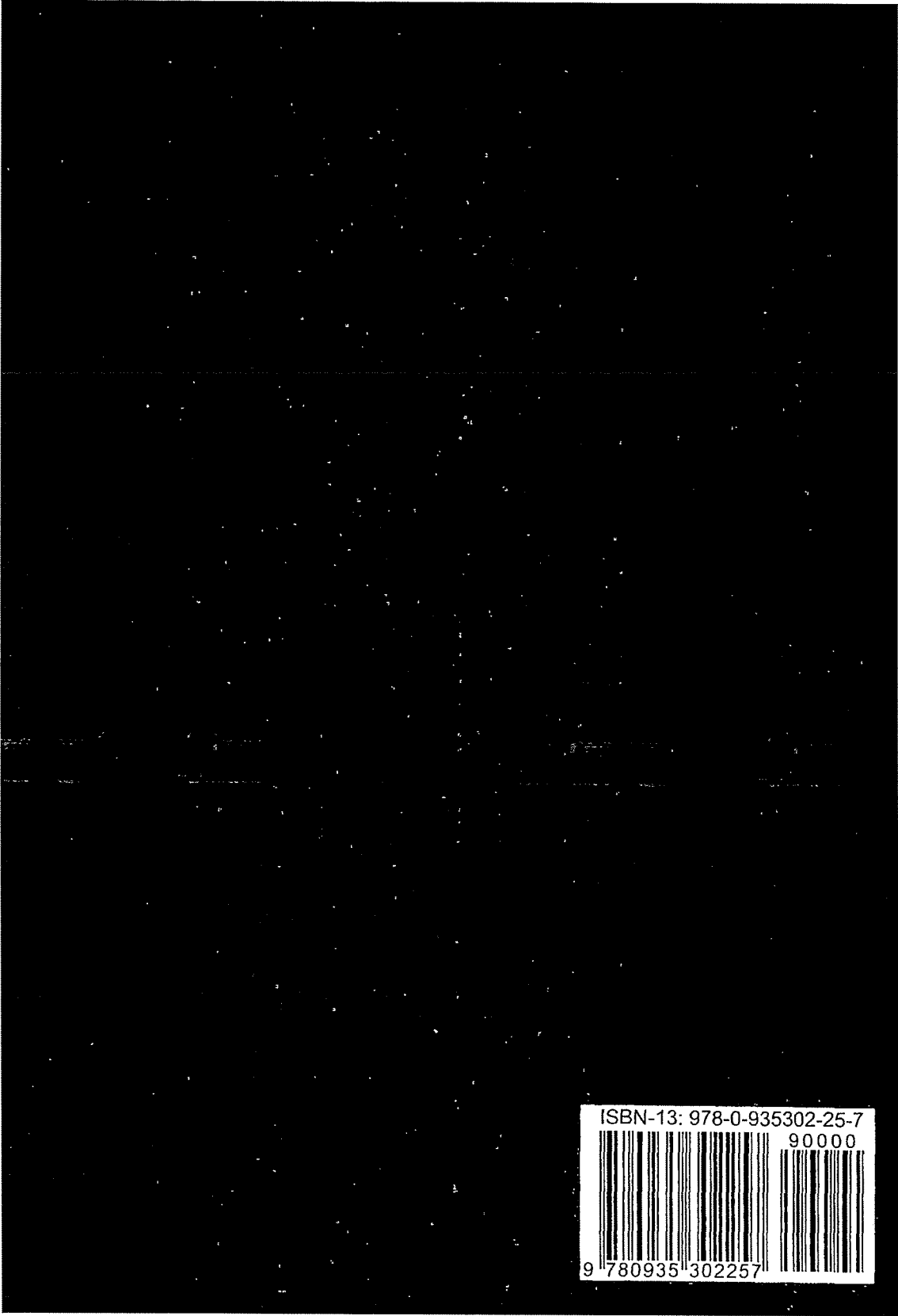
- Structural equation modeling, 13.17, 15.3
- Student outcomes, 13.9
- Target domain**, 13.3
- Test batteries, 12.18
- Test content, 3.6, 7.3-7.4, 8.1
- Test design, 3.15, 7.3
- Test developer responsibilities, see "Publisher materials/responsibilities"
- Test development, 3.1-3.27, 4.19, 6.4, 7.4, 7.7, 7.10, 9.6-9.7, 9.9, 10.1-10.7, 14.1
 - Accommodations for individuals with disabilities, 10.1
 - Comparability of multiple-language versions, 9.9
 - Cut scores, 4.19
 - Definition of domain, 3.2
 - Definition of objective, 14.1
 - Documentation of procedures used to modify test, 10.5
 - Effects of disabilities on test performance, 10.2
 - Effects of modifications for individuals with disabilities, 10.7
 - Empirical procedures to establish time limits for modified forms, 10.6
 - Item selection, 3.6
 - Linguistic or reading level, 7.7
 - Linguistically diverse subgroups, 9.6
 - Pilot testing of modifications for individuals with disabilities, 10.3
 - Rationale for modifications, 10.4
 - Response formats, 3.6
 - Scale development procedures, 6.4
 - Scoring procedures, 3.6
 - Sensitive or offensive content, 7.4
 - Test administration procedures, 3.6
 - Testing outcomes for examinee subgroups, 7.10
 - Translations from one language to another, 9.7
- Test difficulty, 3.3
- Test directions, 3.15
- Test forms, 3.16, 4.10-4.15, 6.5, 7.2, 8.3, 9.4, 9.9, 10.1-10.8, 10.10-10.11, 13.6, 13.17-13.18, 14.17
 - Adapted version in secondary language, 9.4
 - Alternate forms, 4.11, 7.2, 8.3, 14.17
 - Computer administered, 13.18
 - Equated forms, 4.11, 4.13, 6.5, 14.17
 - Interchangeability, 4.10
 - Mixing and distributing for equating studies, 4.12
 - Modifications for individuals with disabilities, 10.1-10.8, 10.10-10.11
 - Multimedia, 13.18
 - Multiple-language versions, 8.3, 9.9
 - Multiple versions from rearrangement of items, 4.15
 - Score equivalence, 4.10-4.11
 - Short form, 3.16
- Test framework, 3.2
- Test information functions, 2.11
- Test interpretation, 2.2-2.3, 7.12, 12.1-12.5, 12.14-12.16, 12.19-12.20, 13.4, 13.12-13.13, 15.4
 - Observed, 2.3
- Test items, 3.6
 - Content quality, 3.6
 - Sensitivity to gender and cultural issues, 3.6
- Test modifications, 2.18, 3.26, 5.1-5.3, 8.3, 9.4-9.5, 9.11, 10.1-10.8, 10.11, 11.23
 - Accommodations for individuals with disabilities, 10.11, 11.23
 - Appropriate for individual test taker, 10.10
 - Documentation, 5.2
 - Documentation of procedures used to modify test, 10.5
 - Effects on resulting scores, 10.7
 - Flagged scores, 9.5, 10.11
 - Individuals with disabilities, 10.2-10.3
 - Interpreters, 9.11
 - Linguistic modifications, 9.4-9.5, 11.23
 - Pilot testing for appropriateness and feasibility, 10.3
 - Psychometric expertise, 10.2
 - Requesting and receiving accommodations, 5.3, 8.3, 10.1-10.2, 10.8
 - Score comparability, 10.4
 - Time limits, 10.6
- Test purpose, see "Purpose of test"
- Test revisions, 3.25-3.26, 4.16
- Test score interpretation, see "Score interpretation"
- Test security, 5.6-5.7, 11.7, 12.11, 13.11
- Test selection, 7.9, 7.11, 10.8, 12.2-12.3, 12.5, 12.6, 12.13, 13.12
 - Addressing complex diagnoses, 12.5
 - Biases, 12.2
 - Culture, 12.3
 - Differential diagnosis, 12.6
 - Language and physical requirements, 12.3
 - Modified forms, 10.8
 - Norms, 12.3
 - Rationale, 12.13
 - Test user qualifications, 12.5, 13.12
 - Validity for population of test taker, 12.3
 - Vested interest, 12.2
- Test settings, 12.8, 13.11
- Test specifications, 3.2-3.5, 3.7, 3.11, 3.14-3.17, 4.16, 6.4, 7.9
 - Changes from one version to subsequent version, 4.16
 - Characteristics, 7.9
 - Consequences, 7.9
 - Definition of content of test, 3.3
 - Definition of domain, 3.14, 3.17
 - Development process, 3.3

INDEX

- Directions to test takers, 3.3
- Information to policy makers, 7.9
- Item and section arrangement, 3.3
- Item formats, 3.3
- Procedures for test administration and scoring, 3.3
- Proposed number of items, 3.3
- Psychometric properties of items, 3.3
- Rationale, 3.3
- Short form, 3.16
- Testing time, 3.3
- Test takers with disabilities, see "Testing individuals with disabilities"
- Test-taking behavior, 12.14
 - Fatigue, 12.14
 - Motivation, 12.14
 - Rapport, 12.14
 - Responses, 12.14
- Test taking strategies, 8.2, 11.13, 15.7, 15.9
 - Negative impact in mandated testing programs, 15.7, 15.9
- Test use, 1.19, 1.21, 1.23, 6.9, 6.15, 7.9-7.11, 9.5-9.6, 10.5, 10.8, 10.11, 11.2-11.3, 14.4-14.5, 14.7, 14.9, 15.10-15.11
 - Consequences, 7.9
 - Employment selection or promotion, 14.9
 - Flagged scores, 9.5, 10.11
 - Job classification decisions, 14.7
 - Justification for testing program, 1.23, 15.10-15.11
 - Linguistically diverse subgroups, 9.5-9.6
 - Studies, 6.9, 14.4-14.5
- Test use rationale, 1.8, 1.11, 12.13
- Test user responsibilities, see "User responsibilities"
- Testing environment, 5.4, 12.12
 - Optimal, 12.12
 - Realistic, 12.12
- Testing for diagnosis, 12.6-12.7
- Testing individuals with disabilities, 10.1-10.12, 11.23
 - Avoiding construct irrelevant variance, 10.1
 - Diagnostic purposes, 10.12
 - Flagged test score, 10.11
 - Functioning relative to general population, 10.9, 11.23
 - Functioning relative to individuals with same level of disability, 10.9
 - Intervention purposes, 10.12
 - Maintaining all feasible standardized features, 10.10
 - Modifications adopted, 10.10
 - Multiple sources of information required, 10.12
 - Not sole indicator of test taker's functioning, 10.12
 - Normative data, 10.9
 - Research of effects of disabilities on test performance, 10.2
- Testing irregularities, 8.10-8.12, 11.11
 - Challenges, 11.11
- Testing policy, 8.2
- Testing programs, 2.18, 2.20, 3.1, 4.17, 8.10-8.13, 9.3, 11.12, 11.20, 13.1-13.19, 15.1, 15.13
- Theoretical foundations of test, 12.18
- Time limits for tests, 3.18, 8.2, 10.6
 - Extensions for modified forms, 10.6
- Translations of a test, 9.7
- Unstructured response format, 12.12
- Use of test scores, 1.1, 1.2, 1.3, 1.4, 7.10-7.11, 8.2, 11.2, 13.1, 13.9, 15.7
 - Cautions about unsupported interpretations, 1.3
 - Decision making for educational placement, 13.9
 - Evidence to justify new use, 1.4, 11.2
 - Mean test score differences between relevant subgroups, 7.10-7.11
- User responsibilities, 1.1, 1.4, 3.24, 4.5, 4.7-4.8, 5.2, 5.7, 5.10, 7.10, 8.7, 9.10, 10.1, 11.1-11.24, 12.1, 12.4-12.5, 12.8-12.9, 12.11-12.12, 13.1, 13.3, 13.10-13.11, 13.19, 15.7, 15.11-15.12
 - Adequate training of supervised test administrators and scorers, 12.8, 13.10
 - Awareness of legal constraints, 11.1, 12.11
 - Consideration of collateral information for test interpretation, 11.20
 - Evaluation of computer-generated interpretations, 11.21
 - Formulate policy for release of aggregated data, 11.17, 13.19
 - General language proficiency of examinee, 9.10, 11.22
 - Identify individuals needing special accommodations, 11.23
 - Informed about purposes and administration of test, 11.5
 - Instructions to individuals who interpret test scores, 12.9, 13.10
 - Interpretive material for local release, 5.10, 11.17-11.18, 13.19, 15.11
 - Justification for use of test, 11.4
 - Minimize or avoid misinterpretations of scores, 11.15, 15.11
 - Monitor impact of mandated testing programs, 13.1, 15.7
 - Monitor scoring accuracy, 11.10
 - Obtain evidence of reliability and validity for new purposes, 11.2
 - Prevent negative consequences, 11.15
 - Professional competence, 12.1, 12.5
 - Professional judgment, 11.1
 - Protect privacy of examinees and institutions, 11.14
 - Protect security of tests, 5.7, 8.7, 11.7-11.9, 12.11, 13.11

INDEX

- Rationale for change in test format or administration, 11.19
- Rationale for intended uses, 11.4-11.5
- Review evidence for using tests in combination, 12.4
- Score reporting, 11.6
- Study and evaluate materials, 11.1
- Test taking strategies, 11.13
- User qualifications, 11.3
- Uses with groups not specified by developer, 7.10
- Verify appropriateness of interpretations, 11.16, 15.11-15.12
- Validation, content-related evidence, 1.6-1.7, 14.8-14.11**
- Validation, criterion-related evidence, 1.15-1.21, 12.17, 14.3
 - Assumptions, 1.21
 - Concurrent study, 1.15
 - Criterion performance, 1.15
 - Criterion relevance, 1.16
 - Differential prediction for groups, 1.19
 - Ethical and legal constraints, 1.19
 - Generalization, 1.20
 - Judgments regarding methodological choices, 1.21
 - Meta-analytic evidence, 1.20-1.21
 - Multiple predictors, 1.17
 - Prediction, 1.17, 14.3
 - Predictive study, 1.15
 - Statistical analysis, 1.17-1.18
 - Technical feasibility, 14.3
 - Test-criterion relationships, 1.16, 1.20
 - Use of test scores, 1.16
- Validation, general issues, 1.1-1.6, 1.13-1.14, 1.22-1.24, 14.1
 - Construct-irrelevant components, 1.24
 - Construct underrepresentation, 1.24
 - Data collection conditions, 1.13
 - Evidence for expected outcome, 1.22
 - Group differences, 1.24
 - Indirect benefit rationale, 1.23
 - Interpretation of test scores, 1.24
 - Objective for employment test, 14.1
 - Statistical analysis, 1.13
 - Testing conditions, 1.13
- Validation procedures, 1.6
- Validation sample, 1.5
- Validity, 1.1-1.24, 3.19, 3.25, 5.12, 6.12, 7.1-7.2, 8.7, 8.11, 9.1-9.2, 9.7, 9.9, 10.1, 10.4-10.5, 10.7, 11.1-11.2, 11.19, 11.22, 12.3-12.6, 12.13, 13.2, 13.7, 13.9, 13.11-13.12, 13.16, 13.18, 14.13, 15.1
 - Changes likely from modifications for individuals with disabilities, 10.5
 - Computer-administered tests, 13.18
 - Computer-generated interpretations, 6.12
 - Construct-irrelevant variance, 1.14
 - Convergent evidence, 1.14
 - Discriminant evidence, 1.14
 - Effects of time passage, 13.16
 - Empirical evidence, 1.8
 - Evidence based on response processes, 1.8
 - Internal consistency evidence, 1.11
 - Interrelationships of scores, 1.11, 1.12
 - Language differences, 9.1
 - Linguistic subgroup validity evidence, 9.2, 11.22
 - Modifications for test takers with disabilities, 10.4
 - Multiple predictors, 13.7, 14.13, 15.1
 - Multiple-purpose tests, 13.2
 - Of a diagnosis, 12.6-12.7
 - Placement or promotion decisions, 13.9
 - Profile interpretation, 1.12
 - Reported for level of aggregation, 5.12
 - Score interpretation rationale, 1.8, 1.11
 - Scores from combination of tests, 12.4-12.5
 - Subgroups, 7.1-7.2
 - Subscore interpretation, 1.12
 - Test comparability, 9.9
 - Test security, 8.7, 13.11
 - Test use rationale, 1.11
 - Testing individuals with disabilities, 10.1
 - Theoretical evidence, 1.8
 - Translations of a test, 9.7
 - Usefulness of modified tests, 10.7
- Validity generalization, 1.20
- Vested interest, 12.2
- Waiver of access, 8.9
- Weighted scoring, 14.16



ISBN-13: 978-0-935302-25-7
90000
9 780935 302257

EXHIBIT Z

Case No. 1:14-cv-00857-TSC-DAR



PLANET DEPOS[®]
We Make It Happen >> Anywhere[™]

CONFIDENTIAL

Transcript of **James R. Fruchterman**

Date: September 8, 2015

Case: American Educational Research Assoc., Inc., et al -v-
Public.Resource.Org., Inc.

Planet Depos
Phone: 888-433-3767
Fax: 888-503-3767
Email: transcripts@planetdepos.com
Internet: www.planetdepos.com

Court Reporting | Videography | Videoconferencing | Interpretation | Transcription

JA2437

1
2
3
4
5
6
7
8
9

10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA

AMERICAN EDUCATIONAL RESEARCH
ASSOCIATION, INC., ET AL.,
PLAINTIFF,
vs. No. 1:14-CV-00857-TSC-DAR
PUBLIC.RESOURCE.ORG, INC.,
DEFENDANT.

VIDEOTAPED DEPOSITION OF
JAMES R. FRUCHTERMAN
CONFIDENTIAL
Tuesday, September 8, 2015

Reported By:
KATHLEEN WILKINS, CSR #10068, RPR-RMR-CRR-CCRR-CLR

1	much narrower area, and I'd say the	01:34:21
2	representations in our seven-point digital rights	01:34:25
3	management plan were the primary mechanism that we	01:34:32
4	dealt with that particular concern of the	01:34:37
5	publishing industry.	01:34:40
6	BY MR. HUDIS:	01:34:41
7	Q. Okay. The last sentence on that page,	01:34:46
8	page 15 of Exhibit 55, it says:	01:34:49
9	"With the extensive input	01:34:51
10	from consumers, authors,	01:34:54
11	publishers and leading	01:34:56
12	organizations, we have created a	01:34:57
13	model for Bookshare that can be	01:34:59
14	supported by a broad array of	01:35:01
15	interests."	01:35:04
16	What model is this passage talking	01:35:05
17	about?	01:35:08
18	MR. KAPLAN: Objection. Lacks	01:35:09
19	foundation.	01:35:10
20	THE WITNESS: The Bookshare operational	01:35:14
21	model.	01:35:17
22	BY MR. HUDIS:	01:35:21
23	Q. How would you describe the Bookshare	01:35:21
24	operational model?	01:35:22
25	A. A package of technologies and policies	01:35:24

1	and legal agreements and product features and -- I	01:35:27
2	mean, you know, it's a -- these things combined	01:35:33
3	create a service that delivers a value to people	01:35:38
4	with disabilities in a way that gets support from	01:35:46
5	these different stakeholders.	01:35:48
6	Q. Including the publishing industry?	01:35:53
7	A. Yes.	01:35:55
8	Q. Could we turn to page 16 of Exhibit 55.	01:35:57
9	Under copyright information, it says:	01:36:00
10	"Bookshare is an online	01:36:02
11	library that provides accessible	01:36:04
12	eBooks to people with print	01:36:06
13	disabilities. Bookshare meets the	01:36:07
14	requirements of the Chafee	01:36:09
15	Amendment which permits an	01:36:09
16	authorized entity like Benetech to	01:36:12
17	make books available to people	01:36:14
18	with print disabilities provided	01:36:16
19	that copies may not be reproduced	01:36:17
20	or distributed in a format other	01:36:19
21	than a specialized format	01:36:21
22	exclusively for use by blind or	01:36:23
23	other persons with disabilities.	01:36:25
24	Must bear a notice that any	01:36:27
25	further reproduction or	01:36:32

1	distribution in a format other	01:36:33
2	than a specialized format is an	01:36:35
3	infringement. Must include a	01:36:37
4	copyright notice identifying the	01:36:39
5	copyright owner and the date of	01:36:43
6	the original publication.	01:36:45
7	'Specialized formats' means	01:36:46
8	Braille, audio or digital text	01:36:50
9	which is exclusively intended for	01:36:53
10	use by blind or other persons with	01:36:54
11	disabilities."	01:36:56
12	All right. So I've read this passage,	01:36:59
13	Mr. Fruchterman.	01:37:01
14	A. Right.	01:37:01
15	Q. Does this accurately describe the	01:37:01
16	overall way that Benetech makes reading materials	01:37:03
17	available to its members?	01:37:07
18	MR. KAPLAN: Objection. Vague.	01:37:08
19	Misleading.	01:37:09
20	THE WITNESS: I think that these bullet	01:37:14
21	points that you just read recapitulate the	01:37:16
22	provisions of the Chafee Amendment, which is the	01:37:19
23	primary copyright exception that we use for making	01:37:23
24	copyright material to people with qualifying	01:37:26
25	disabilities inside the United States.	01:37:28

1	BY MR. HUDIS:	01:37:31
2	Q. If we could go to page 17 of Exhibit 55.	01:37:31
3	What is the purpose of this page on	01:37:36
4	Bookshare's web site?	01:37:38
5	MR. KAPLAN: Objection. Vague. Lacks	01:37:40
6	foundation.	01:37:41
7	THE WITNESS: This is part of our,	01:37:44
8	essentially, frequently asked questions, and it's	01:37:45
9	entitled "Digital Millennium Copyright Act."	01:37:49
10	And so as a -- and I'm not a lawyer, but	01:37:54
11	my understanding is is someone who provides access	01:37:58
12	to copyrighted material online, we are required to	01:38:02
13	have a DMCA agent to accept notices that there is	01:38:06
14	content on our web site that infringes the	01:38:12
15	copyright of others.	01:38:14
16	We frequently get DMCA notices from	01:38:17
17	authors or their agents or publishers saying, We	01:38:23
18	searched the web. This copyright work is on your	01:38:26
19	web site. Take it down.	01:38:29
20	And this is both explaining the DMCA	01:38:30
21	notice process at some level, as well as the, more	01:38:36
22	or less, if you don't know what the Chafee	01:38:40
23	Amendment is, you should look it up because we're	01:38:42
24	allowed to have it.	01:38:47
25	But I'm summarizing this in very direct	01:38:48

172

1	terms, because it's very rare for someone to issue	01:38:54
2	us a DMCA notice that results in us actually	01:38:56
3	taking down the work because it's usually legally	01:39:01
4	permitted under the copyright amendment.	01:39:04
5	BY MR. HUDIS:	01:39:05
6	Q. The Chafee Amendment to the copyright?	01:39:06
7	A. The Chafee Amendment. Or often a	01:39:07
8	license from the author's publisher who gave us	01:39:10
9	the content, but the author and their agent	01:39:12
10	weren't aware this was one of the nice things that	01:39:14
11	their publisher did for their entire catalog of	01:39:17
12	books, not just that author.	01:39:21
13	Q. Mr. Fruchterman, could we turn to page	01:39:23
14	18 of Exhibit 55.	01:39:25
15	Is this text on page 18 Bookshare's	01:39:34
16	digital rights plan -- digital rights management	01:39:40
17	plan?	01:39:46
18	A. This is the current or, let's just say,	01:39:46
19	last month's current -- but I don't believe it's	01:39:49
20	changed since last month -- version of our	01:39:51
21	seven-point digital rights management plan that we	01:39:53
22	have discussed earlier.	01:39:56
23	Q. And what was the purpose of Bookshare	01:39:58
24	implementing this DRM plan?	01:39:59
25	MR. KAPLAN: Objection. Vague. Lacks	01:40:04

1	foundation.	01:40:05
2	THE WITNESS: I would say that the	01:40:11
3	purpose of this was to represent to the	01:40:12
4	intellectual property industry, especially	01:40:17
5	publishers, that we were intending to follow the	01:40:19
6	law when it came to use of these materials. So it	01:40:22
7	was created for that original conversation we had	01:40:25
8	with the publishing industry quite a number of	01:40:27
9	years ago.	01:40:31
10	BY MR. HUDIS:	01:40:31
11	Q. And when you say "these materials,"	01:40:32
12	that's the copyrighted materials on the Bookshare	01:40:34
13	web site?	01:40:36
14	MR. KAPLAN: Objection. Misstates	01:40:39
15	testimony.	01:40:40
16	THE WITNESS: Yes.	01:40:42
17	BY MR. HUDIS:	01:40:43
18	Q. Could we turn to page 19.	01:40:43
19	A. Mh-hmm.	01:40:46
20	Q. What's the purpose of this sign-up page?	01:40:46
21	That's page 19 of Exhibit 55.	01:40:52
22	MR. KAPLAN: Objection. Vague. Lacks	01:40:54
23	foundation.	01:40:55
24	THE WITNESS: This is a screen shot that	01:41:15
25	appears to be of the individual sign-up for	01:41:16

1	Bookshare that is collecting data about a	01:41:22
2	potential user in order to start the process of	01:41:24
3	becoming a Bookshare member.	01:41:29
4	BY MR. HUDIS:	01:41:32
5	Q. And at the bottom it says -- it has a	01:41:32
6	check box, and then you would sign your name or	01:41:34
7	its equivalent.	01:41:36
8	Do you see at the bottom?	01:41:38
9	A. Yes.	01:41:39
10	Q. And by doing so you're agreeing to the	01:41:39
11	terms and conditions of the Bookshare web site.	01:41:42
12	Do you see that?	01:41:44
13	MR. KAPLAN: Objection. Is the -- the	01:41:45
14	question is whether or not he sees that check box?	01:41:49
15	MR. HUDIS: Counsel, good.	01:41:53
16	Q. Is the purpose of this check box to have	01:41:55
17	the user acknowledge that he or she is agreeing to	01:42:02
18	the terms and conditions of the Bookshare web	01:42:04
19	site?	01:42:07
20	MR. KAPLAN: Objection. Vague. Lacks	01:42:08
21	foundation.	01:42:09
22	MR. HUDIS: Thank you, Counsel.	01:42:10
23	THE WITNESS: Yes. I believe that that	01:42:13
24	check box and the filling in of your name	01:42:14
25	indicates that you're agreeing to the terms and	01:42:17

1	conditions of our -- of our -- of our agreement,	01:42:19
2	of our Bookshare individual membership agreement.	01:42:22
3	BY MR. HUDIS:	01:42:25
4	Q. And if you could turn to page 20 of	01:42:25
5	Exhibit 55. Are those the terms and conditions of	01:42:26
6	the -- of the Bookshare web site?	01:42:31
7	A. It appears to be our standard Bookshare	01:42:34
8	membership agreement of a recent date.	01:42:38
9	MR. HUDIS: Counsel, same request. Can	01:42:48
10	we stipulate this is a business record of	01:42:49
11	Benetech?	01:42:53
12	MR. KAPLAN: Subject to your	01:42:58
13	representation that this is -- each page	01:42:59
14	represents a complete Snagit screen shot of a	01:43:04
15	particular web site or web page of the Benetech	01:43:07
16	web site, I believe so.	01:43:12
17	But can we go off the record for just a	01:43:15
18	second?	01:43:18
19	MR. HUDIS: Yes. I consent. We can go	01:43:19
20	off the record.	01:43:20
21	THE VIDEOGRAPHER: Okay. Going off the	01:43:21
22	record at 1:43.	01:43:21
23	(Discussion held off record.)	01:43:40
24	THE VIDEOGRAPHER: Back on the record at	01:43:50
25	1:43.	01:43:51

1 MR. KAPLAN: So subject to Counsel's 01:43:53
2 representation regarding the contents of this 01:43:55
3 exhibit, we stipulate to its authenticity as 01:43:57
4 select web pages from the Benetech web site. 01:44:03
5 MR. HUDIS: All right. Now, that's the 01:44:06
6 authenticity. What about business record? That 01:44:07
7 was what I was concerned about. You stipulated to 01:44:09
8 the authenticity. We do have -- I do -- 01:44:14
9 MR. KAPLAN: You want a stipulation that 01:44:20
10 the statements in here are not hearsay for the 01:44:22
11 purpose of -- 01:44:25
12 MR. HUDIS: For what they contain. 01:44:27
13 MR. KAPLAN: I don't believe we can 01:44:41
14 stipulate that -- to that because, as far as I 01:44:42
15 know, we don't represent Benetech. 01:44:45
16 BY MR. HUDIS: 01:44:49
17 Q. All right. So if you could -- if, 01:44:49
18 Mr. Fruchterman, you could put Exhibit 55 back in 01:44:52
19 front of you. 01:44:56
20 A. Yes. 01:44:58
21 Q. All right. So the pages on Exhibit 55, 01:44:58
22 I'm going to represent to you that they are Snagit 01:45:02
23 screen shots of the Bookshare web site. 01:45:04
24 So my question is are these pages items 01:45:08
25 of data compilations made by Benetech? 01:45:11

1	MR. HUDIS: No. I'm having problems	03:03:29
2	with --	03:03:30
3	MR. KAPLAN: Scrolling.	03:03:31
4	MR. HUDIS: -- what was put down as your	03:03:34
5	answer.	03:03:34
6	(Record read by the reporter	03:03:41
7	as follows:	03:03:41
8	ANSWER: As before, I would	03:03:41
9	change numbers that are based on	03:03:41
10	the date of this declaration.)	03:03:41
11	BY MR. HUDIS:	03:03:41
12	Q. So which numbers would you change?	03:03:42
13	MR. KAPLAN: Objection. Vague.	03:03:44
14	THE WITNESS: Yeah. Paragraph 1 -- 2 --	03:03:45
15	sorry, paragraph 2, I cite how many users, how	03:03:48
16	many books, what our monthly capacity is. I would	03:03:52
17	update those to current figures.	03:03:58
18	BY MR. HUDIS:	03:03:59
19	Q. So it would be more?	03:03:59
20	A. Yes.	03:04:01
21	MR. KAPLAN: Description.	03:04:01
22	THE WITNESS: Sorry.	03:04:02
23	That's it.	03:04:26
24	BY MR. HUDIS:	03:04:27
25	Q. Are paragraphs 4 through 12 of	03:04:27

1	Exhibit 60 still today an accurate description of	03:04:32
2	Bookshare's seven-point digital rights management	03:04:36
3	plan?	03:04:40
4	MR. KAPLAN: Objection. Vague.	03:04:40
5	THE WITNESS: Yes.	03:04:47
6	BY MR. HUDIS:	03:04:49
7	Q. If we could turn to paragraph 1, page 1,	03:04:50
8	of Exhibit 60. You say:	03:04:53
9	"Based upon my experience	03:04:55
10	with the Bookshare online library	03:04:57
11	for people with print	03:04:59
12	disabilities, I believe that the	03:05:00
13	risk of online piracy or	03:05:02
14	unauthorized copying and	03:05:04
15	distribution of works made fully	03:05:05
16	available to individuals" --	03:05:07
17	"individuals with print	03:05:09
18	disabilities through the	03:05:12
19	HathiTrust is minimal."	03:05:13
20	What was the basis for this statement	03:05:17
21	that you made in paragraph 1?	03:05:19
22	MR. KAPLAN: Objection. Confusing. The	03:05:23
23	document speaks for itself. Vague.	03:05:25
24	THE WITNESS: My declaration explains	03:05:29
25	why, at length.	03:05:31

1	BY MR. HUDIS:	03:05:34
2	Q. Why is there no discussion of the	03:05:34
3	HathiTrust security measures in this declaration	03:05:36
4	of Exhibit 60?	03:05:39
5	MR. KAPLAN: Objection. Argumentative.	03:05:44
6	Vague.	03:05:47
7	And I will instruct the witness not to	03:05:49
8	answer to the extent that it calls for privileged	03:05:50
9	communications or information protected by Rule 26	03:05:54
10	of the Federal Rules of Civil Procedure.	03:05:58
11	BY MR. HUDIS:	03:06:01
12	Q. Mr. Fruchterman, first of all, will you	03:06:01
13	adhere to counsel's instructions?	03:06:05
14	A. Yes.	03:06:07
15	MR. KAPLAN: First --	03:06:07
16	BY MR. HUDIS:	03:06:08
17	Q. And can you --	03:06:08
18	MR. KAPLAN: Yeah. Okay.	03:06:09
19	BY MR. HUDIS:	03:06:10
20	Q. And can you answer my question without	03:06:10
21	revealing the substance of attorney-client	03:06:13
22	communications?	03:06:16
23	A. No.	03:06:18
24	Q. In making the statement "I believe that	03:06:25
25	the risk of online piracy or unauthorized copying	03:06:29

1	and distribution of works made fully available to	03:06:33
2	individuals with print disabilities through the	03:06:35
3	HathiTrust is minimal," did you review the	03:06:39
4	security measures on the HathiTrust web site?	03:06:42
5	MR. KAPLAN: Objection. Vague.	03:06:46
6	THE WITNESS: Not beyond previously	03:06:50
7	discussed.	03:06:52
8	BY MR. HUDIS:	03:06:55
9	Q. Mr. Fruchterman, do you recall what the	03:07:11
10	outcome was in the HathiTrust litigation?	03:07:13
11	MR. KAPLAN: Objection. Vague. Calls	03:07:17
12	for a legal conclusion. Lacks foundation.	03:07:18
13	THE WITNESS: I do.	03:07:20
14	MR. KAPLAN: I'm sorry. Scratch the	03:07:20
15	last objection.	03:07:22
16	THE WITNESS: I do.	03:07:24
17	BY MR. HUDIS:	03:07:26
18	Q. All right. What -- and what was -- what	03:07:27
19	is your understanding of the outcome of the	03:07:28
20	HathiTrust litigation?	03:07:30
21	A. That the motion for summary judgment by	03:07:34
22	the defendants was granted by the district court	03:07:38
23	judgment and upheld in an appellate court	03:07:46
24	decision.	03:07:51
25	Q. And did you -- did you review the	03:07:57

1	district court's opinion after it was issued?	03:07:59
2	A. I did.	03:08:04
3	(Whereupon, Deposition Exhibit 61 was	03:08:34
4	marked for identification.)	03:08:34
5	BY MR. HUDIS:	03:08:40
6	Q. Mr. Fruchterman, I'd like you to turn to	03:08:41
7	page 4 of what's now been marked as Exhibit 61.	03:08:45
8	It is the district court's opinion in the Authors	03:08:50
9	Guild, Inc. versus HathiTrust, et al., reported at	03:08:52
10	<u>902 F.Supp.2d 445</u> and the date of the decision is	03:08:58
11	October 10, 2012.	03:09:02
12	MR. KAPLAN: Counsel, it's a Westlaw	03:09:04
13	printout.	03:09:06
14	MR. HUDIS: Yes.	03:09:07
15	MR. KAPLAN: Including Westlaw's	03:09:07
16	commentary and descriptions and additional	03:09:08
17	material that was not contained in the original	03:09:10
18	decision.	03:09:12
19	MR. HUDIS: Noted.	03:09:14
20	Q. Mr. Fruchterman, could you please turn	03:09:17
21	to page 4 of the document.	03:09:19
22	A. Yes.	03:09:22
23	Q. And it says, under "Background,"	03:09:23
24	"Defendants"-- are you with me?	03:09:26
25	A. Yes.	03:09:29

1	Q. All right.	03:09:30
2	"Defendants have entered into	03:09:31
3	agreements with Google Inc. that	03:09:31
4	allow Google to create digital	03:09:33
5	copies of works in the	03:09:35
6	universities' libraries in	03:09:38
7	exchange for which Google provides	03:09:39
8	digital copies to defendants, the	03:09:41
9	mass digitization product or MDP."	03:09:44
10	Was that your understanding of how the	03:09:47
11	HathiTrust library worked?	03:09:49
12	MR. KAPLAN: Objection. Vague.	03:09:53
13	Confusing.	03:09:56
14	THE WITNESS: Yes. Generally.	03:09:59
15	BY MR. HUDIS:	03:10:01
16	Q. All right. If we could turn to page 5	03:10:01
17	of Exhibit 61. At the top left-hand corner, it	03:10:03
18	says:	03:10:10
19	"After digitization, Google	03:10:10
20	retains a copy of the digital book	03:10:11
21	that is available through Google	03:10:13
22	Books, an online system through	03:10:14
23	which Google users can search the	03:10:17
24	content and view snippets of the	03:10:19
25	books. Google also provides a	03:10:21

1	away?	05:30:02
2	Q. Soon.	05:30:02
3	A. Or are these the only two I need to have	05:30:03
4	out now?	05:30:05
5	Q. Those are the only two you have to have	05:30:07
6	out now.	05:30:09
7	A. Okay. I have those two documents in	05:30:18
8	front of me, Exhibit 55 and 60.	05:30:19
9	Q. Okay. So I would like to focus your	05:30:25
10	attention on -- in the supplemental declaration,	05:30:25
11	Exhibit 60, to pages 2 and 3, where you talk about	05:30:28
12	the digital rights management plan.	05:30:33
13	A. Yes.	05:30:37
14	Q. Okay. And similarly, an explanation of	05:30:38
15	the DRM plan on page 18 of Exhibit 55. And that's	05:30:43
16	the Bookshare web site.	05:31:04
17	A. Okay.	05:31:13
18	Q. During your review of Public.Resource's	05:31:13
19	web site, how did their web site compare with the	05:31:16
20	Bookshare web site in terms of employing a digital	05:31:20
21	rights management or DRM plan to protect the	05:31:23
22	digital copies of standards posted on	05:31:27
23	Public.Resource's web site from unauthorized	05:31:30
24	copying?	05:31:35
25	MR. KAPLAN: Objection. Vague. Calls	05:31:35

1	for a legal conclusion. Confusing.	05:31:36
2	THE WITNESS: I didn't find a DRM plan	05:31:43
3	in evidence on the Public.Resource.Org site.	05:31:45
4	MR. HUDIS: I'd like to take a break for	05:31:52
5	five minutes.	05:33:18
6	THE VIDEOGRAPHER: Going off the record	05:33:19
7	at 5:33.	05:33:20
8	(Whereupon, a recess was taken.)	05:33:27
9	THE VIDEOGRAPHER: Back on the record at	05:39:35
10	5:39.	05:39:37
11	BY MR. HUDIS:	05:39:38
12	Q. Mr. Fruchterman, when you examined	05:39:40
13	Public.Resource's web site, you noticed a number	05:39:44
14	of standards that were hosted on that web site?	05:39:48
15	A. Correct.	05:39:58
16	MR. KAPLAN: Objection. Vague. Asked	05:39:59
17	and answered.	05:40:00
18	BY MR. HUDIS:	05:40:00
19	Q. Did you notice any restrictions on the	05:40:01
20	ability of an Internet user to copy any of the	05:40:02
21	standards that you saw on Public.Resource's web	05:40:11
22	site?	05:40:13
23	MR. KAPLAN: Objection. Vague.	05:40:14
24	THE WITNESS: No.	05:40:17
25		

328

1	BY MR. HUDIS:	05:40:17
2	Q. Did you notice any restrictions on the	05:40:19
3	ability of an Internet user to download any of the	05:40:20
4	standards hosted on the Public.Resource's web	05:40:28
5	site?	05:40:31
6	MR. KAPLAN: Objection. Vague.	05:40:31
7	THE WITNESS: No.	05:40:32
8	BY MR. HUDIS:	05:40:32
9	Q. Did you notice any restrictions on the	05:40:32
10	ability of an Internet user to print any of the	05:40:35
11	standards hosted on the Public.Resource web site?	05:40:37
12	MR. KAPLAN: Objection. Vague.	05:40:40
13	THE WITNESS: No.	05:40:41
14	MR. HUDIS: Thank you, Mr. Fruchterman.	05:40:43
15	That's all I have.	05:40:43
16	THE WITNESS: Okay. Thank you.	05:40:46
17	MR. KAPLAN: I have no questions at this	05:40:52
18	time.	05:40:53
19	THE WITNESS: Okay. Oh, that's right.	05:40:53
20	You get a chance, huh.	05:40:54
21	THE VIDEOGRAPHER: This marks the end of	05:40:56
22	the deposition of James Fruchterman. Going off	05:40:56
23	the record at 5:41.	05:40:59
24	(Whereupon, the deposition concluded	05:41:00
25	at 5:41 p.m.)	05:41:00

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

CERTIFICATE OF REPORTER

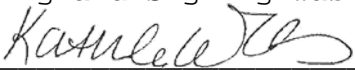
I, Kathleen A. Wilkins, Certified Shorthand Reporter licensed in the State of California, License No. 10068, hereby certify that the deponent was by me first duly sworn, and the foregoing testimony was reported by me and was thereafter transcribed with computer-aided transcription; that the foregoing is a full, complete, and true record of proceedings.

I further certify that I am not of counsel or attorney for either or any of the parties in the foregoing proceeding and caption named or in any way interested in the outcome of the cause in said caption.

The dismantling, unsealing, or unbinding of the original transcript will render the reporter's certificates null and void.

In witness whereof, I have hereunto set my hand this day:

- Reading and Signing was requested.
- Reading and Signing was waived.
- Reading and Signing was not requested.



KATHLEEN A. WILKINS

CSR 10068, RPR-RMR-CRR-CCRR-CLR

EXHIBIT II

Case No. 1:14-cv-00857-TSC-DAR

```
archive — ssh — 80x24
butler@vm-home0:~$ date
Tue Nov 25 00:01:58 UTC 2014
butler@vm-home0:~$ sql "select identifier,downloads from item_stats where identi
fier='gov.law.aera.standards.1999'"
  identifier          | downloads
-----+-----
 gov.law.aera.standards.1999 |      1290
butler@vm-home0:~$
```

Witness Butler
Pl. 11
Def. 11
Exhibit _____
Consisting of _____ Pages
Date 12-2-14
CINDY TUGAW, CSR 4805

EXHIBIT JJ

Case No. 1:14-cv-00857-TSC-DAR

From: John S. Neikirk
To: 'carl@media.org'
Sent: 12/16/2013 11:08:00 PM
Subject: copyright infringement

Dear Carl Malamud,

I am writing on behalf of the American Educational Research Association in regard to a copyrighted work, Standards for Educational and Psychological Testing. Without permission, the volume is posted at the following url:

<https://law.resource.org/pub/us/cfr/ibr/001/aera.standards.1999.pdf>

Please remove this unlawful posting immediately.

Cordially,

John Neikirk
Director of Publications
American Educational Research Association
1430 K Street, NW, Suite 1200
Washington, DC 20005
202.238.3238
jneikirk@aera.net



JA2461

AERA_APA_NCME_0005129

EXHIBIT KK

Case No. 1:14-cv-00857-TSC-DAR



PUBLIC.RESOURCE.ORG ~ A Nonprofit Corporation

Open Source America's Operating System

"It's Not Just A Good Idea—It's The Law!"



December 19, 2013

John Neikirk
 Director of Publications
 American Educational Research Association
 1430 K Street, NW, Suite 1200
 Washington, DC 2005

Dear Mr. Neikerk:

I am receipt of your communication of December 16 regarding the publication of the AERA publication, "Standard for Educational and Psychological Testing" (1999) at <https://law.resource.org/pub/us/cfr/ibr/001/aera.standards.1999.pdf>. We are responsible for uploading this document. In addition, you will find this document at <https://archive.org/details/gov.law.aera.standards.1999>.

The 1999 Edition of "Standard for Educational and Psychological Testing" was Incorporated by Reference by the Department of Education, Office of Postsecondary Education, at 34 CFR 668.148(a)(1)(iv). Incorporation by reference is not a casual affair and requires a carefully followed procedure by the governmental agency and the explicit approval of the Director of the Office of the Federal Register.

As this standard has been incorporated into law, the standard contained in this document is the law of the United States, and people in the United States are compelled to obey it. Long-standing precedent of the United States Supreme Court holds that copyright claims cannot prevent citizens from reading and speaking the law. See *Wheaton v. Peters*, 33 U.S. 591 (1834); *Banks v. Manchester*, 128 U.S. 244 (1888).

While the standards drafted by the American Educational Research Association, were entitled to copyright protection when issued, once they were incorporated into regulations these standards became the law, and thus have entered the public domain. Chief Judge Edith H. Jones of the 5th Circuit expressed this principle clearly in her opinion in *Veeck v. Southern Building Code Congress*, which concerned a model building code incorporated in the law of two Texas towns:

"The issue in this en banc case is the extent to which a private organization may assert copyright protection for its model codes, after the models have been adopted by a legislative body and become "the law." Specifically, may a code-

writing organization prevent a website operator from posting the text of a model code where the code is identified simply as the building code of a city that enacted the model code as law? Our short answer is that as law, the model codes enter the public domain and are not subject to the copyright holder's exclusive prerogatives. As model codes, however, the organization's works retain their protected status." 293 F.3d 791 (5th Cir. 2002) (en banc).

As you can see by looking at the document in question, a cover sheet has been prepended clearly spelling out the section of the Code of Federal Regulations that has incorporated by reference this document into law. Please note that we were careful to only publish the specific document incorporated by law. As the 1999 Edition is the one required by law and as it has been duly incorporated into law, we respectfully decline to remove this document and respectfully decline to request permission.

Sincerely yours,



Digitally signed by Carl
Malamud
DN: cn=Carl Malamud,
o=Public.Resource.Org
,ou,
email=carl@media.org,
c=US
Date: 2013.12.19
10:34:04 -08'00'

Carl Malamud

EXHIBIT MM

Case No. 1:14-cv-00857-TSC-DAR



PUBLIC.RESOURCE.ORG ~ A Nonprofit Corporation

Public Works for a Better Government

Re: American Educational Research Association., Inc. et al v.
Public.Resource.Org, Inc., No. 1:14-cv-00857

This memorandum is in reference to the lawsuit named above, which concerns the document entitled "Standards for Educational and Psychological Testing" which was duly incorporated by the Office of the Federal Register into the Code of Federal Regulations and specifically in response to the stated intention to file a preliminary injunction motion.

Public.Resource.Org believes firmly that because the document in question has been explicitly incorporated into federal law, it has the right to post it on its website, and that it will prevail in this case. Public Resource also believes that this case deserves the court's fullest consideration, without a rush to reach an interim ruling in the absence of a full record.

In order to focus this case on developing an appropriate record for a decision on the merits, Public.Resource.Org has voluntarily removed the document in question from the websites under its control and has removed the document from all public access on the Internet Archive.

Until the conclusion of trial on the merits in this case, Public.Resource.Org will keep the document in question off of the websites under its control and will not disseminate the document, in whole or in part, including any revisions, and will maintain the status on the Internet Archive to prevent any public access to the document from the Archive's websites.

Public.Resource.Org believes that this action obviates any need for plaintiffs to rush the court to a judgment on a partial record of their own selection without a full opportunity for all parties to develop the facts and issues of the case for trial.

Carl Malamud, President and Founder

June 12, 2014
Date



**UNITED STATES DISTRICT COURT
 FOR THE DISTRICT OF COLUMBIA**

AMERICAN EDUCATIONAL RESEARCH)	
ASSOCIATION, INC., AMERICAN)	
PSYCHOLOGICAL ASSOCIATION, INC.,)	
and NATIONAL COUNCIL ON)	
MEASUREMENT IN EDUCATION, INC.,)	Civil Action No. 1:14-cv-00857-CRC
)	
Plaintiffs,)	DECLARATION OF MARIANNE
)	ERNESTO IN SUPPORT OF
v.)	PLAINTIFFS' MOTION FOR
)	SUMMARY JUDGMENT AND ENTRY
PUBLIC.RESOURCE.ORG, INC.,)	OF A PERMANENT INJUNCTION
)	
Defendant.)	
)	

I, MARIANNE ERNESTO, declare:

1. I am the Director, Testing and Assessment, at the American Psychological Association, Inc. ("APA"). I have been employed with the APA since May 2001. I submit this Declaration in support of the motion of the American Educational Research Association, Inc. ("AERA"), the APA, and the National Council on Measurement in Education, Inc. ("NCME") (collectively, "Plaintiffs" or the "Sponsoring Organizations") for a summary judgment and the entry of a permanent injunction.

2. In my role as Director, Testing and Assessment, I serve as APA's primary authority on all matters that relate to testing and assessment. This subject matter includes educational testing, clinical assessment, forensic testing and employment testing. I advocate on behalf of APA in matters involving federal or state legislative, regulatory or other policy issues concerning testing and assessment. I coordinate APA's involvement in testing issues in matters such as governance, executive boards, and managerial bodies. I also manage APA's responses to internal, public, member and media inquiries regarding testing issues in a manner that is consistent with the *Standards for Educational and Psychological Testing* (the "Standards"). I

advise, counsel and oversee the activities of the APA's Science Directorate (and in particular its Office of Testing and Assessment) on policy and governance issues related to testing and assessments. I further serve as staff liaison to the APA's Committee on Psychological Tests and Assessment ("CPTA"). Since 2001, I have served as APA's primary contact for information concerning the availability and interpretation of the Standards published in 1999, and more recently I have done so regarding the updated Standards published in 2014.

3. APA is a District of Columbia not-for-profit corporation.

4. APA is the largest scientific and professional organization representing psychology in the United States. APA is the world's largest association of psychologists and counts a vast number of researchers, educators, clinicians, consultants and students among its members. APA's mission is to advance the creation, communication and application of psychological knowledge to benefit society and improve people's lives.

5. In 1954, APA prepared and published the "Technical Recommendations for Psychological Tests and Diagnostic Techniques." It is my understanding that in 1955 AERA and NCME prepared and published a companion document entitled, "Technical Recommendations for Achievement Tests."

6. Subsequently, a joint committee of the three organizations modified, revised and consolidated the two documents into the first Joint Standards. Beginning with the 1966 revision, the three organizations (AERA, APA and NCME) collaborated in developing the "Joint Standards" (or simply, the "Standards"). Each subsequent revision of the Standards has been careful to cite the previous Standards and note that it is a revision and update of that document.

7. Beginning in the mid-1950s, AERA, APA, and NCME formed and periodically reconstituted a committee of experts in psychological and educational assessment, charged with

the initial development of the Technical Recommendations and then each subsequent revision of the (renamed) Standards. These committees were formed by the Plaintiffs' Presidents (or their designees), who would meet and jointly agree on the membership. Often a chair or co-chairs of these committees were selected by joint agreement. Beginning with the 1966 version of the Standards, this committee became referred to as the "Joint Committee."

8. Financial and operational oversight for the Standards' revisions, promotion, distribution, and for the sale of the 1999 and 2014 Standards has been undertaken by a periodically reconstituted Management Committee, comprised of designees of the three Sponsoring Organizations.

9. All members of the Joint Committee(s) and the Management Committee(s) are *unpaid* volunteers. The expenses associated with the ongoing development and publication of the Standards include travel and lodging expenses (for the Joint Committee and Management Committee members), support staff time, printing and shipment of bound volumes, and advertising costs.

10. Many different fields of endeavor rely on assessments. The Sponsoring Organizations have ensured that the range of these fields of endeavor is represented in the Joint Committee's membership – *e.g.*, admissions, achievement, clinical counseling, educational, licensing-credentialing, employment, policy, and program evaluation. Similarly, the Joint Committee's members, who are *unpaid volunteers*, represent expertise across major functional assessment areas – *e.g.*, validity, equating, reliability, test development, scoring, reporting, interpretation, and large scale interpolation.

11. From the time of their initial creation to the present, the preparation of and periodic revisions to the Standards entail intensive labor and considerable cross-disciplinary

expertise. Each time the Standards are revised, the Sponsoring Organizations select and arrange for meetings of the leading authorities in psychological and educational assessments (known as the Joint Committee). During these meetings, certain Standards are combined, pared down, and/or augmented, others are deleted altogether, and some are created as whole new individual Standards. The 1999 version of the Standards is nearly 200 pages and took more than five years to complete.

12. The Standards, however, are not simply intended for members of the Sponsoring Organizations, AERA, APA, and NCME. The intended audience of the Standards is broad and cuts across audiences with varying backgrounds and different training. For example, the Standards also are intended to guide test developers, sponsors, publishers, and users by providing criteria for the evaluation of tests, testing practices, and the effects of test use. Test user standards refer to those standards that help test users decide how to choose certain tests, interpret scores, or make decisions based on tests results. Test users include clinical or industrial psychologists, research directors, school psychologists, counselors, employment supervisors, teachers, and various administrators who select or interpret tests for their organizations. There is no mechanism, however, to enforce compliance with the Standards on the part of the test developer or test user. The Standards, moreover, do not attempt to provide psychometric answers to policy or legal questions.

13. The Standards apply broadly to a wide range of standardized instruments and procedures that sample an individual's behavior, including tests, assessments, inventories, scales, and other testing vehicles. The Standards apply equally to standardized multiple-choice tests, performance assessments (including tests comprised of only open-ended essays), and hands-on assessments or simulations. The main exceptions are that the Standards do not apply to

unstandardized questionnaires (*e.g.*, unstructured behavioral checklists or observational forms), teacher-made tests, and subjective decision processes (*e.g.*, a teacher's evaluation of students' classroom participation over the course of a semester).

14. The Standards have been used to develop testing guidelines for such activities as college admissions, personnel selection, test translations, test user qualifications, and computer-based testing. The Standards also have been widely cited to address technical, professional, and operational norms for all forms of assessments that are professionally developed and used in a variety of settings. The Standards additionally provide a valuable public service to state and federal governments as they voluntarily choose to use them. For instance, each testing company, when submitting proposals for testing administration, instead of relying on a patchwork of local, or even individual and proprietary, testing design and implementation criteria, may rely instead on the Sponsoring Organizations' Standards to afford the best guidance for testing and assessment practices.

15. The Standards were not created or updated to serve as a legally binding document, in response to an expressed governmental or regulatory need, nor in response to any legislative action or judicial decision. However, the Standards have been cited in judicial decisions related to the proper use and evidence for assessment, as well as by state and federal legislators. These citations in judicial decisions and during legislative deliberations occurred without any lobbying by the Plaintiffs.

16. During the discovery phase of this litigation, APA located in its archives correspondence relating to APA's support for proposed legislation sought to be introduced in 2001 by Senator Paul Wellstone (D-MN) on Fairness and Accuracy in High Stakes Educational Decisions for Students - a suggested amendment to the Elementary and Secondary Education Act

(“No Child Left Behind Act”) 147 Cong. Rec. S. 4,644 (daily ed. May 9, 2001).

17. Accompanying this Declaration as Exhibit NN is a true copy of a signed correspondence between Ellen Garrison Ph.D. and Patricia Kobor of APA and Ms. Jill Morningstar, Legislative Assistant to U.S. Senator Paul Wellstone dated April 7, 2000, marked as Exhibit 1109 during my deposition.

18. Accompanying this Declaration as Exhibit OO is a true copy of an unsigned correspondence between Ellen Garrison Ph.D. and Patricia Kobor of APA and Ms. Jill Morningstar, Legislative Assistant to U.S. Senator Paul Wellstone dated April 7, 2000, marked as Exhibit 1110 during my deposition.

19. Accompanying this Declaration as Exhibit PP is a true copy of a signed correspondence between Patricia Kobor and Ellen Garrison, Ph.D. of APA and Ms. Jill Morningstar, Legislative Assistant to U.S. Senator Paul Wellstone dated April 13, 2000, marked as Exhibit 1111 during my deposition.

20. Accompanying this Declaration as Exhibit QQ is a true copy of an unsigned correspondence between Raymond D. Fowler, Ph.D. of APA and an unnamed Senator dated May 7, 2001, marked as Exhibit 1114 during my deposition.

21. Accompanying this Declaration as Exhibit RR is a true copy of an unsigned correspondence between L. Michael Honaker, Ph.D. of APA and an unnamed Senator dated March 6, 2001, marked as Exhibit 1115 during my deposition.

22. Accompanying this Declaration as Exhibit SS is a true copy of a document containing “Highlights of APA’s Involvement in Educational Testing Provisions of the ‘No Child Left Behind Act’” that also contains an unsigned correspondence to an unnamed Senator dated May 7, 2001, marked as Exhibit 1116 during my deposition.

23. As noted above, many of these letters are unsigned and are not printed on APA letterhead. Therefore, in accordance with APA practices and protocols, it is likely that the unsigned letters (not printed on letterhead) were internal discussion drafts that were never sent.

24. Regarding the signed letters that were printed on APA letterhead, they relate to Senator Wellstone's proposed legislation that tests and assessments administered by the states are of high quality and used appropriately for the benefit of test administrators and test takers. These are goals that are consistent with APA policy as then reflected in the 1999 Standards. Even though Senator Wellstone's amendments sought, in part, to mandate states' compliance with the Standards, none of the Sponsoring Organizations actively advocated for this – and in any event Senator Wellstone's proposed amendment including this language was never enacted into law. Accompanying this Declaration as Exhibit TT is a true copy of 20 U.S.C. § 6301, which is the current version of the legislation Senator Wellstone sought to amend.

25. APA's search of its records did not disclose any further communications with Congress relating to the Standards and, to the best of APA's knowledge, it has not engaged in communications with Congress regarding citation of the Standards in legislation since 2001.

26. APA has not solicited any government agency to incorporate the Standards into the Code of Federal Regulations or other rules of Federal or State agencies.

27. Rather, in the policymaking arena, APA believes the Standards should be treated as guidelines informing the enactment of legislation and regulations consistent with best practices in the development and use of tests – to insure that they are valid, reliable and fair.

28. Plaintiffs promote and sell copies of the Standards via referrals to the AERA website, at annual meetings, in public offerings to students, and to educational institution faculty. Advertisements promoting the Standards have appeared in meeting brochures, in scholarly

journals, and in the hallways at professional meetings. Accompanying this Declaration as Exhibit UU is a true copy of an advertisement for the 1999 Standards that appeared in the December 1999 issue of APA's Journal of Educational Psychology.

29. Distribution of the Standards is closely monitored by the Sponsoring Organizations. AERA, the designated publisher of the Standards, sometimes does provide promotional complementary print copies to students or professors. Except for these few complementary print copies, however, the Standards are not given away for free; and certainly they are not made available to the public by any of the three organizations for anyone to copy free of charge.

30. To date, the Sponsoring Organizations have never posted, or authorized the posting of, a digitized copy of the 1999 Standards on any publicly accessible website.

31. The Sponsoring Organizations do not keep any of the revenues generated from the sales of the Standards. Rather, the income from these sales is used by the Sponsoring Organizations to offset their development and production costs and to generate funds for subsequent revisions. This allows the Sponsoring Organizations to develop up-to-date, high quality Standards that otherwise would not be developed due to the time and effort that goes into producing them.

32. Without receiving at least some moderate income from the sales of the Standards to offset their production costs and to allow for further revisions, it is very likely that the Sponsoring Organizations would no longer undertake to periodically update them, and it is unknown who else would.

33. Due to the relative minor portion of the membership of APA who devote their careers to testing and assessment, it is highly unlikely that the members of APA will vote for a

dues increase to fund future Standards revision efforts if Public Resource successfully defends this case and is allowed to post the Standards online for the public to download or print for free. As a result, the Sponsoring Organizations would likely abandon their practice of periodically updating the Standards.

34. The Joint Committee that authored the 1999 Standards comprised 16 members. Except for Manfred Meier (who could not be located, nor could his heirs), work made-for-hire letters were signed by 13 Joint Committee Members, and posthumous assignments were signed by the heirs of 2 deceased Joint Committee Members, vesting ownership of the copyright to the 1999 Standards in the Sponsoring Organizations. Accompanying this Declaration as Exhibits VV-HHH are the 13 work made-for-hire letters signed by Eva Baker, Lloyd Bond, Daniel Goh, Bert Green, Edward Haertel, Jo-Ida Hansen, Suzanne Lane, Sharon Johnson-Lewis, Joseph Matarazzo, Pamela Moss, Esteban Olmedo, Diana Pullin, and Paul Sackett, marked as Exhibits 1065, 1069, 1071, 1072, 1075, 1078, 1082, 1085, 1086, 1089, 1090, 1091, and 1094 during my deposition. Accompanying this Declaration as Exhibits III and JJJ are the posthumous assignments signed by the heirs of Leonard S. Feldt and Charlie Spielberger, marked as exhibits 1070 and 1097 during my deposition.

35. Public Resource posted Plaintiffs' 1999 Standards to its website and the Internet Archive website without the permission or authorization of any of the Sponsoring Organizations.

36. Past harm from Public Resource's infringing activities includes misuse of Plaintiffs' intellectual property without permission.

37. Should Public Resource's infringement be allowed to continue, the harm to the Sponsoring Organizations, and public at large who rely on the preparation and administration of valid, fair and reliable tests, includes: (i) uncontrolled publication of the 1999 Standards without

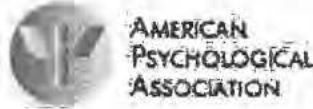
any notice that those guidelines have been replaced by the 2014 Standards; (ii) future unquantifiable loss of revenue from sales of authorized copies of the 1999 Standards (with proper notice that they are no longer the current version) and the 2014 Standards; and (iii) lack of funding for future revisions of the 2014 Standards and beyond.

Dated: December 8, 2015

Marianne Ernesto
Marianne Ernesto

EXHIBIT VV

Case No. 1:14-cv-00857-TSC-DAR



April 21, 2014

Dear Member of the Joint Committee to Revised the 1985 *Standards for Educational and Psychological Testing*:

In 1998 and 1999, we, the American Educational Research Association ("AERA"), the American Psychological Association ("APA") and the National Council on Measurement in Education ("NCME"), (collectively, the "Publishers"), commissioned you and other leaders in the educational research, psychology and educational testing fields to contribute information, materials and acumen to the collective work entitled *Standards for Educational and Psychological Testing*" (the "Standards"). Since its publication in 1999, the Publishers had and still have the right to use the Standards in print, electronic format any and all other formats then known, now known or hereafter to become known. We are confirming in this letter that you accepted this assignment subject to the following terms and conditions:

1. You delivered manuscript(s) or review(s) of manuscript(s) within the time period established by Publishers. You were reimbursed for reasonable expenses in connection with your work on the Standards, if approved by the Publishers in advance, upon your submission of receipts. Your name appeared in the Preface of the Standards, showing that you were one of its contributors, and you received a free copy of the Standards upon its publication.
2. You acknowledge that the Standards, and all contributions that you made toward completion and publication of the Standards, was and still is considered a "work made for hire" within the meaning of the United States copyright laws, and that the Publishers own all right, title and interest in and to the copyright in the Standards. To the extent that the Standards were not or are not deemed to be a "work made for hire" you hereby assign nunc pro tunc (now for then) to the Publishers all right, title and interest in and to the Standards.
3. Accordingly, the Publishers may also use the Standards for any and all uses and products, in any and all formats now known or hereafter to become known, including but not limited to print, recorded on hard storage media (e.g., CDs, DVDs, etc.), the Internet and online services.
4. You have granted the Publishers the right to use your name in the Standards, in advertising and promotion related to the Standards, and in any and all ancillary products related to the Standards regardless of the formats in which such use occurs.
5. Your contributions to the Standards were wholly original material not published elsewhere (except for material in the public domain or used with the permission of the owner), did not and does not infringe any copyright, and did not and does not constitute a defamation, or invasion of the right of privacy or publicity, or infringement of any other kind, of any third party.
6. It is specifically understood and intended that you are an independent contractor, and nothing herein is intended or shall be deemed to make you an employee of any of the Publishers.

If the foregoing accurately sets forth our understanding, please sign and date, then scan and return this letter via e-mail attachment to Marianne Ernesto, Director, Testing and Assessment, American Psychological Association, at mernesto@apa.org.

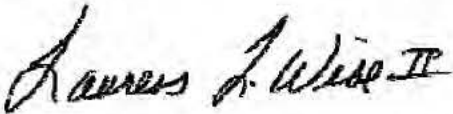
Sincerely,



AMERICAN EDUCATIONAL RESEARCH ASSOCIATION



AMERICAN PSYCHOLOGICAL ASSOCIATION



NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

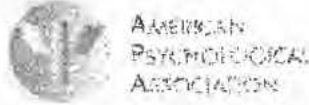
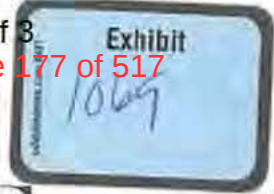
Accepted and Agreed:



Date: 4/24/14

EXHIBIT WW

Case No. 1:14-cv-00857-TSC-DAR



April 21, 2014

Dear Member of the Joint Committee to Revised the 1985 *Standards for Educational and Psychological Testing*:

In 1998 and 1999, we, the American Educational Research Association ("AERA"), the American Psychological Association ("APA") and the National Council on Measurement in Education ("NCME"), (collectively, the "Publishers"), commissioned you and other leaders in the educational research, psychology and educational testing fields to contribute information, materials and acumen to the collective work entitled *Standards for Educational and Psychological Testing* (the "Standards"). Since its publication in 1999, the Publishers had and still have the right to use the Standards in print, electronic format any and all other formats then known, now known or hereafter to become known. We are confirming in this letter that you accepted this assignment subject to the following terms and conditions:

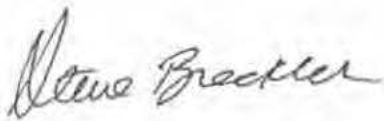
1. You delivered manuscript(s) or review(s) of manuscript(s) within the time period established by Publishers. You were reimbursed for reasonable expenses in connection with your work on the Standards, if approved by the Publishers in advance, upon your submission of receipts. Your name appeared in the Preface of the Standards, showing that you were one of its contributors, and you received a free copy of the Standards upon its publication.
2. You acknowledge that the Standards, and all contributions that you made toward completion and publication of the Standards, was and still is considered a "work made for hire" within the meaning of the United States copyright laws, and that the Publishers own all right, title and interest in and to the copyright in the Standards. To the extent that the Standards were not or are not deemed to be a "work made for hire" you hereby assign nunc pro tunc (now for then) to the Publishers all right, title and interest in and to the Standards.
3. Accordingly, the Publishers may also use the Standards for any and all uses and products, in any and all formats now known or hereafter to become known, including but not limited to print, recorded on hard storage media (e.g., CDs, DVDs, etc.), the Internet and online services.
4. You have granted the Publishers the right to use your name in the Standards, in advertising and promotion related to the Standards, and in any and all ancillary products related to the Standards regardless of the formats in which such use occurs.
5. Your contributions to the Standards were wholly original material not published elsewhere (except for material in the public domain or used with the permission of the owner), did not and does not infringe any copyright, and did not and does not constitute a defamation, or invasion of the right of privacy or publicity, or infringement of any other kind, of any third party.
6. It is specifically understood and intended that you are an independent contractor, and nothing herein is intended or shall be deemed to make you an employee of any of the Publishers.

If the foregoing accurately sets forth our understanding, please sign and date, then scan and return this letter via e-mail attachment to Marianne Ernesto, Director, Testing and Assessment, American Psychological Association, at mernesto@apa.org.

Sincerely,



AMERICAN EDUCATIONAL RESEARCH ASSOCIATION

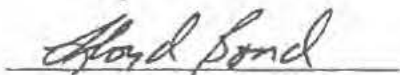


AMERICAN PSYCHOLOGICAL ASSOCIATION



NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

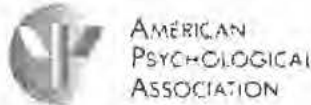
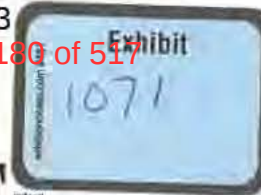
Accepted and Agreed:



Date: 10 September 2014

EXHIBIT XX

Case No. 1:14-cv-00857-TSC-DAR



April 21, 2014

Dear Member of the Joint Committee to Revised the 1985 *Standards for Educational and Psychological Testing*:

In 1998 and 1999, we, the American Educational Research Association ("AERA"), the American Psychological Association ("APA") and the National Council on Measurement in Education ("NCME"), (collectively, the "Publishers"), commissioned you and other leaders in the educational research, psychology and educational testing fields to contribute information, materials and acumen to the collective work entitled *Standards for Educational and Psychological Testing* (the "Standards"). Since its publication in 1999, the Publishers had and still have the right to use the Standards in print, electronic format any and all other formats then known, now known or hereafter to become known. We are confirming in this letter that you accepted this assignment subject to the following terms and conditions:

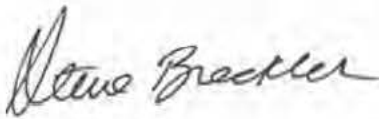
1. You delivered manuscript(s) or review(s) of manuscript(s) within the time period established by Publishers. You were reimbursed for reasonable expenses in connection with your work on the Standards, if approved by the Publishers in advance, upon your submission of receipts. Your name appeared in the Preface of the Standards, showing that you were one of its contributors, and you received a free copy of the Standards upon its publication.
2. You acknowledge that the Standards, and all contributions that you made toward completion and publication of the Standards, was and still is considered a "work made for hire" within the meaning of the United States copyright laws, and that the Publishers own all right, title and interest in and to the copyright in the Standards. To the extent that the Standards were not or are not deemed to be a "work made for hire" you hereby assign nunc pro tunc (now for then) to the Publishers all right, title and interest in and to the Standards.
3. Accordingly, the Publishers may also use the Standards for any and all uses and products, in any and all formats now known or hereafter to become known, including but not limited to print, recorded on hard storage media (e.g., CDs, DVDs, etc.), the Internet and online services.
4. You have granted the Publishers the right to use your name in the Standards, in advertising and promotion related to the Standards, and in any and all ancillary products related to the Standards regardless of the formats in which such use occurs.
5. Your contributions to the Standards were wholly original material not published elsewhere (except for material in the public domain or used with the permission of the owner), did not and does not infringe any copyright, and did not and does not constitute a defamation, or invasion of the right of privacy or publicity, or infringement of any other kind, of any third party.
6. It is specifically understood and intended that you are an independent contractor, and nothing herein is intended or shall be deemed to make you an employee of any of the Publishers.

If the foregoing accurately sets forth our understanding, please sign and date, then scan and return this letter via e-mail attachment to Marianne Ernesto, Director, Testing and Assessment, American Psychological Association, at mernesto@apa.org.


Sincerely,



AMERICAN EDUCATIONAL RESEARCH ASSOCIATION

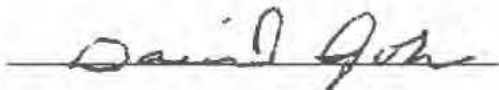


AMERICAN PSYCHOLOGICAL ASSOCIATION



NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

Accepted and Agreed:



Date: 9-8-2014

EXHIBIT YY

Case No. 1:14-cv-00857-TSC-DAR



April 21, 2014

Dear Member of the Joint Committee to Revised the 1985 Standards for Educational and Psychological Testing:

In 1998 and 1999, we, the American Educational Research Association ("AERA"), the American Psychological Association ("APA") and the National Council on Measurement in Education ("NCME"), (collectively, the "Publishers"); commissioned you and other leaders in the educational research, psychology and educational testing fields to contribute information, materials and acumen to the collective work entitled Standards for Educational and Psychological Testing (the "Standards"). Since its publication in 1999, the Publishers had and still have the right to use the Standards in print, electronic format any and all other formats then known, now known or hereafter to become known. We are confirming in this letter that you accepted this assignment subject to the following terms and conditions:

1. You delivered manuscript(s) or review(s) of manuscript(s) within the time period established by Publishers. You were reimbursed for reasonable expenses in connection with your work on the Standards; if approved by the Publishers in advance, upon your submission of receipts. Your name appeared in the Preface of the Standards, showing that you were one of its contributors; and you received a free copy of the Standards upon its publication.
2. You acknowledge that the Standards, and all contributions that you made toward completion and publication of the Standards, was and still is considered a "work made for hire" within the meaning of the United States copyright laws, and that the Publishers own all right, title and interest in and to the copyright in the Standards. To the extent that the Standards were not or are not deemed to be a "work made for hire", you hereby assign, nunc pro tunc (now for then) to the Publishers all right, title and interest in and to the Standards.
3. Accordingly, the Publishers may also use the Standards for any and all uses and products, in any and all formats now known or hereafter to become known, including but not limited to print, recorded on hard storage media (e.g., CDs, DVDs, etc.), the Internet and online services.
4. You have granted the Publishers the right to use your name in the Standards, in advertising and promotion related to the Standards, and in any and all ancillary products related to the Standards regardless of the formats in which such use occurs.
5. Your contributions to the Standards were wholly original material not published elsewhere (except for material in the public domain or used with the permission of the owner), did not and does not infringe any copyright, and did not and does not constitute a defamation, or invasion of the right of privacy or publicity, or infringement of any other kind, of any third party.
6. It is specifically understood and intended that you are an independent contractor, and nothing herein is intended or shall be deemed to make you an employee of any of the Publishers.

If the foregoing accurately sets forth our understanding, please sign and date, then scan and return this letter via e-mail attachment to Marianne Ernesto, Director, Testing and Assessment, American Psychological Association, at mernesto@apa.org

Sincerely,

[Signature]

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION

[Signature]

AMERICAN PSYCHOLOGICAL ASSOCIATION

[Signature]

NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

Accepted and Agreed:

[Signature]
Date: 4/21/2014

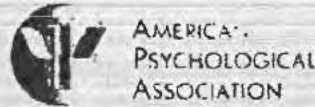
J1A2487

Case 1:14-cv-00857-TSC Document 60-61 Filed 12/21/15 Page 2 of 2

EXHIBIT ZZ

Case No. 1:14-cv-00857-TSC-DAR

Exhibit
1075



April 21, 2014

Dear Member of the Joint Committee to Revised the 1985 *Standards for Educational and Psychological Testing*:

In 1998 and 1999, we, the American Educational Research Association ("AERA"), the American Psychological Association ("APA") and the National Council on Measurement in Education ("NCME"), (collectively, the "Publishers"), commissioned you and other leaders in the educational research, psychology and educational testing fields to contribute information, materials and acumen to the collective work entitled *Standards for Educational and Psychological Testing* (the "Standards"). Since its publication in 1999, the Publishers had and still have the right to use the Standards in print, electronic format and all other formats then known, now known or hereafter to become known. We are confirming in this letter that you accepted this assignment subject to the following terms and conditions:

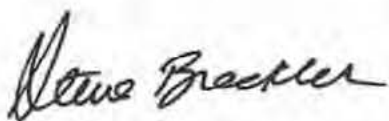
1. You delivered manuscript(s) or review(s) of manuscript(s) within the time period established by Publishers. You were reimbursed for reasonable expenses in connection with your work on the Standards, if approved by the Publishers in advance, upon your submission of receipts. Your name appeared in the Preface of the Standards, showing that you were one of its contributors, and you received a free copy of the Standards upon its publication.
2. You acknowledge that the Standards, and all contributions that you made toward completion and publication of the Standards, was and still is considered a "work made for hire" within the meaning of the United States copyright laws, and that the Publishers own all right, title and interest in and to the copyright in the Standards. To the extent that the Standards were not or are not deemed to be a "work made for hire" you hereby assign nunc pro tunc (now for then) to the Publishers all right, title and interest in and to the Standards.
3. Accordingly, the Publishers may also use the Standards for any and all uses and products, in any and all formats now-known or hereafter to become known, including but not limited to print, recorded on hard storage media (e.g., CDs, DVDs, etc.), the Internet and online services.
4. You have granted the Publishers the right to use your name in the Standards, in advertising and promotion related to the Standards, and in any and all ancillary products related to the Standards regardless of the formats in which such use occurs.
5. Your contributions to the Standards were wholly original material not published elsewhere (except for material in the public domain or used with the permission of the owner), did not and does not infringe any copyright, and did not and does not constitute a defamation, or invasion of the right of privacy or publicity, or infringement of any other kind, of any third party.
6. It is specifically understood and intended that you are an independent contractor, and nothing herein is intended or shall be deemed to make you an employee of any of the Publishers.

If the foregoing accurately sets forth our understanding, please sign and date, then scan and return this letter via e-mail attachment to Marianne Ernesto, Director, Testing and Assessment, American Psychological Association, at mernesto@apa.org.

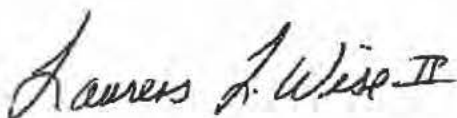
Sincerely,



AMERICAN EDUCATIONAL RESEARCH ASSOCIATION



AMERICAN PSYCHOLOGICAL ASSOCIATION



NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

Accepted and Agreed:

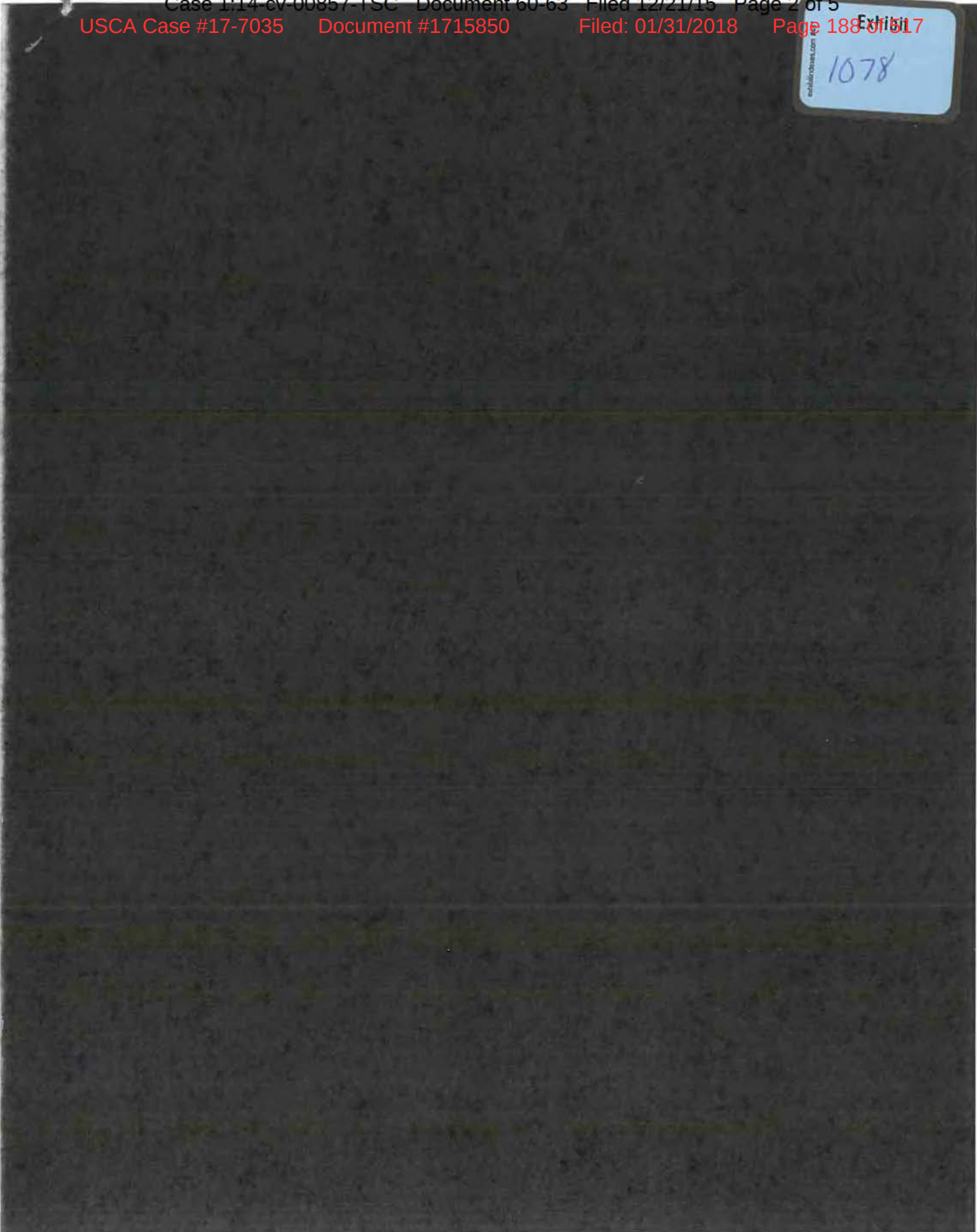


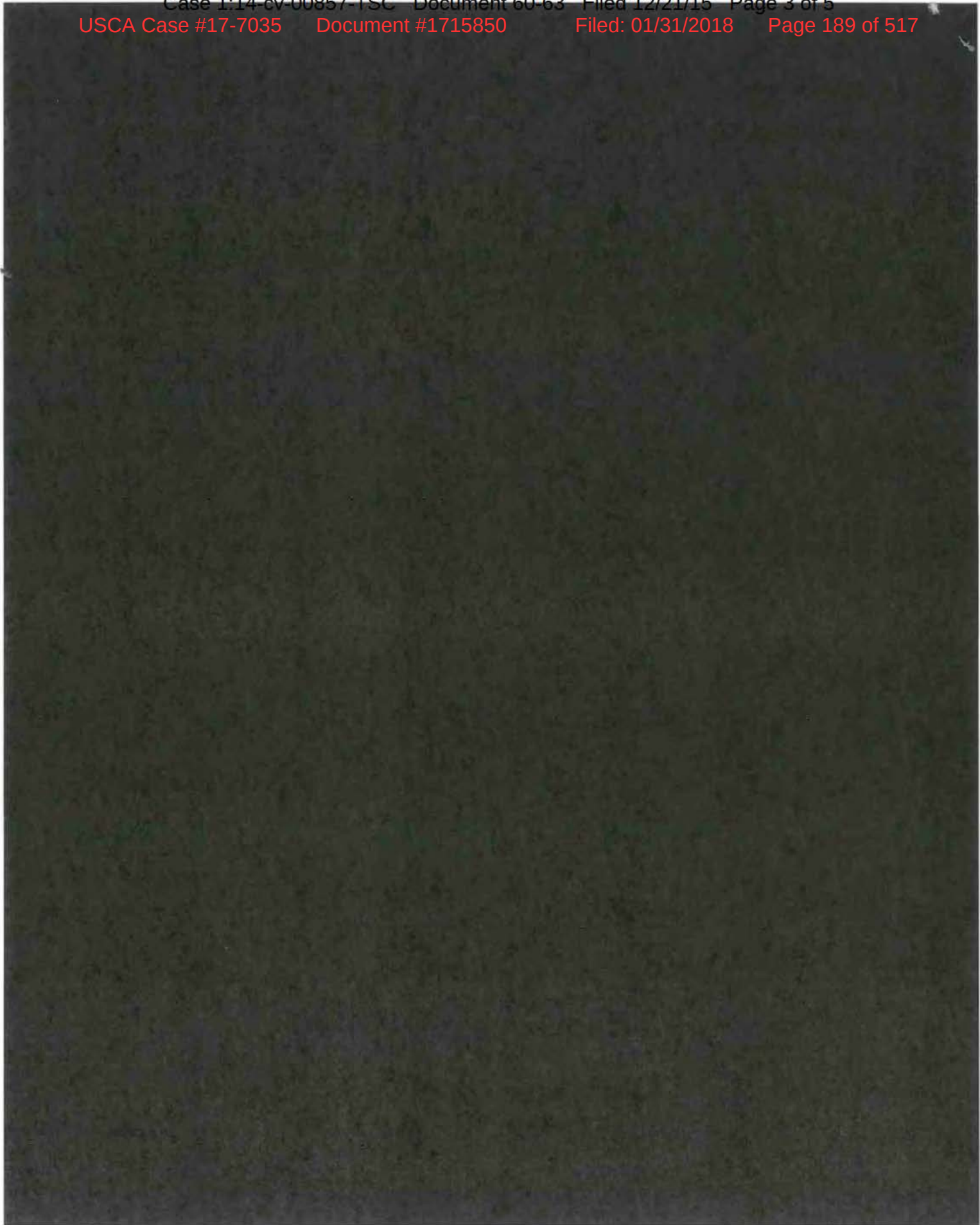
Date: April 21, 2014

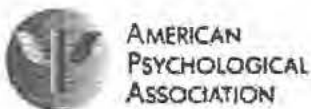
EXHIBIT AAA

Case No. 1:14-cv-00857-TSC-DAR

Exhibit
1078







April 21, 2014

Dear Member of the Joint Committee to Revised the 1985 Standards for Educational and Psychological Testing:

In 1998 and 1999, we, the American Educational Research Association ("AERA"), the American Psychological Association ("APA") and the National Council on Measurement in Education ("NCME"), (collectively, the "Publishers"), commissioned you and other leaders in the educational research, psychology and educational testing fields to contribute information, materials and acumen to the collective work entitled Standards for Educational and Psychological Testing" (the "Standards"). Since its publication in 1999, the Publishers had and still have the right to use the Standards in print, electronic format any and all other formats then known, now known or hereafter to become known. We are confirming in this letter that you accepted this assignment subject to the following terms and conditions:

1. You delivered manuscript(s) or review(s) of manuscript(s) within the time period established by Publishers. You were reimbursed for reasonable expenses in connection with your work on the Standards, if approved by the Publishers in advance, upon your submission of receipts. Your name appeared in the Preface of the Standards, showing that you were one of its contributors, and you received a free copy of the Standards upon its publication.
2. You acknowledge that the Standards, and all contributions that you made toward completion and publication of the Standards, was and still is considered a "work made for hire" within the meaning of the United States copyright laws, and that the Publishers own all right, title and interest in and to the copyright in the Standards. To the extent that the Standards were not or are not deemed to be a "work made for hire" you hereby assign nunc pro tunc (now for then) to the Publishers all right, title and interest in and to the Standards.
3. Accordingly, the Publishers may also use the Standards for any and all uses and products, in any and all formats now known or hereafter to become known, including but not limited to print, recorded on hard storage media (e.g., CDs, DVDs, etc.), the Internet and online services.
4. You have granted the Publishers the right to use your name in the Standards, in advertising and promotion related to the Standards, and in any and all ancillary products related to the Standards regardless of the formats in which such use occurs.
5. Your contributions to the Standards were wholly original material not published elsewhere (except for material in the public domain or used with the permission of the owner), did not and does not infringe any copyright, and did not and does not constitute a defamation, or invasion of the right of privacy or publicity, or infringement of any other kind, of any third party.
6. It is specifically understood and intended that you are an independent contractor, and nothing herein is intended or shall be deemed to make you an employee of any of the Publishers.

Confidential

If the foregoing accurately sets forth our understanding, please sign and date, then scan and return this letter via e-mail attachment to Marianne Ernesto, Director, Testing and Assessment, American Psychological Association, at mernesto@apa.org.

Sincerely,



AMERICAN EDUCATIONAL RESEARCH ASSOCIATION



AMERICAN PSYCHOLOGICAL ASSOCIATION



NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

Accepted and Agreed:

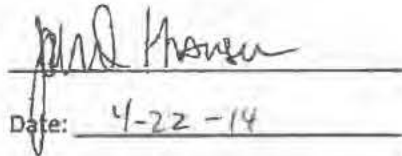
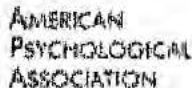

Date: 4-22-14

EXHIBIT BBB

Case No. 1:14-cv-00857-TSC-DAR



April 21, 2014

Dear Member of the Joint Committee to Revised the 1985 *Standards for Educational and Psychological Testing*:

In 1998 and 1999, we, the American Educational Research Association ("AERA"), the American Psychological Association ("APA") and the National Council on Measurement in Education ("NCME"), (collectively, the "Publishers"), commissioned you and other leaders in the educational research, psychology and educational testing fields to contribute information, materials and acumen to the collective work entitled *Standards for Educational and Psychological Testing*" (the "Standards"). Since its publication in 1999, the Publishers had and still have the right to use the Standards in print, electronic format any and all other formats then known, now known or hereafter to become known. We are confirming in this letter that you accepted this assignment subject to the following terms and conditions:

1. You delivered manuscript(s) or review(s) of manuscript(s) within the time period established by Publishers. You were reimbursed for reasonable expenses in connection with your work on the Standards, if approved by the Publishers in advance, upon your submission of receipts. Your name appeared in the Preface of the Standards, showing that you were one of its contributors, and you received a free copy of the Standards upon its publication.
2. You acknowledge that the Standards, and all contributions that you made toward completion and publication of the Standards, was and still is considered a "work made for hire" within the meaning of the United States copyright laws, and that the Publishers own all right, title and interest in and to the copyright in the Standards. To the extent that the Standards were not or are not deemed to be a "work made for hire" you hereby assign nunc pro tunc (now for then) to the Publishers all right, title and interest in and to the Standards.
3. Accordingly, the Publishers may also use the Standards for any and all uses and products, in any and all formats now known or hereafter to become known, including but not limited to print, recorded on hard storage media (e.g., CDs, DVDs, etc.), the Internet and online services.
4. You have granted the Publishers the right to use your name in the Standards, in advertising and promotion related to the Standards, and in any and all ancillary products related to the Standards regardless of the formats in which such use occurs.
5. Your contributions to the Standards were wholly original material not published elsewhere (except for material in the public domain or used with the permission of the owner), did not and does not infringe any copyright, and did not and does not constitute a defamation, or invasion of the right of privacy or publicity, or infringement of any other kind, of any third party.
6. It is specifically understood and intended that you are an independent contractor, and nothing herein is intended or shall be deemed to make you an employee of any of the Publishers.



If the foregoing accurately sets forth our understanding, please sign and date, then scan and return this letter via e-mail attachment to Marianne Ernesto, Director, Testing and Assessment, American Psychological Association, at mernesto@apa.org.

Sincerely,

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION

AMERICAN PSYCHOLOGICAL ASSOCIATION

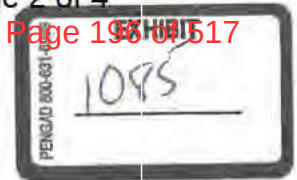
NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

Accepted and Agreed:

Date: April 21, 2017

EXHIBIT CCC

Case No. 1:14-cv-00857-TSC-DAR



April 21, 2014

Dear Member of the Joint Committee to Revised the 1985
Standards for Educational and Psychological Testing:

In 1998 and 1999, we, the American Educational Research Association (“AERA”), the American Psychological Association (“APA”) and the National Council on Measurement in Education (“NCME”), (collectively, the “Publishers”), commissioned you and other leaders in the educational research, psychology and educational testing fields to contribute information, materials and acumen to the collective work entitled *Standards for Educational and Psychological Testing*” (the “Standards”). Since its publication in 1999, the Publishers had and still have the right to use the Standards in print, electronic format any and all other formats then known, now known or hereafter to become known. We are confirming in this letter that you accepted this assignment subject to the following terms and conditions:

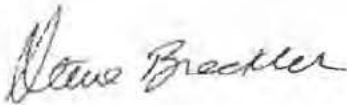
1. You delivered manuscript(s) or review(s) of manuscript(s) within the time period established by Publishers. You were reimbursed for reasonable expenses in connection with your work on the Standards, if approved by the Publishers in advance, upon your submission of receipts. Your name appeared in the Preface of the Standards, showing that you were one of its contributors, and you received a free copy of the Standards upon its publication.

2. You acknowledge that the Standards, and all contributions that you made toward completion and publication of the Standards, was and still is considered a “work made for hire” within the meaning of the United States copyright laws, and that the Publishers own all right, title and interest in and to the copyright in the Standards. To the extent that the Standards were not or are not deemed to be a “work made for hire” you hereby assign nunc pro tunc (now for then) to the Publishers all right, title and interest in and to the Standards.
3. Accordingly, the Publishers may also use the Standards for any and all uses and products, in any and all formats now known or hereafter to become known, including but not limited to print, recorded on hard storage media (e.g., CDs, DVDs, etc.), the Internet and online services.
4. You have granted the Publishers the right to use your name in the Standards, in advertising and promotion related to the Standards, and in any and all ancillary products related to the Standards regardless of the formats in which such use occurs.
5. Your contributions to the Standards were wholly original material not published elsewhere (except for material in the public domain or used with the permission of the owner), did not and does not infringe any copyright, and did not and does not constitute a defamation, or invasion of the right of privacy or publicity, or infringement of any other kind, of any third party.
6. It is specifically understood and intended that you are an independent contractor, and nothing herein is intended or shall be deemed to make you an employee of any of the Publishers. If the foregoing accurately sets forth our understanding, please sign and date, then scan and return this letter via e-mail attachment to Marianne Ernesto, Director, Testing and Assessment, American Psychological Association, at

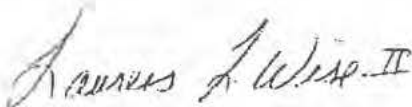
Sincerely,



AMERICAN EDUCATIONAL RESEARCH
ASSOCIATION



AMERICAN PSYCHOLOGICAL ASSOCIATION



NATIONAL COUNCIL ON MEASUREMENT IN
EDUCATION

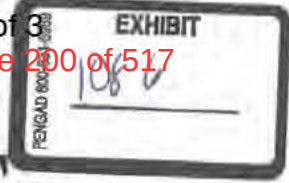
Accepted and Agreed:



Date: 10/16/14

EXHIBIT DDD

Case No. 1:14-cv-00857-TSC-DAR



April 21, 2014

Dear Member of the Joint Committee to Revised the 1985 *Standards for Educational and Psychological Testing*:

In 1998 and 1999, we, the American Educational Research Association ("AERA"), the American Psychological Association ("APA") and the National Council on Measurement in Education ("NCME"), (collectively, the "Publishers"), commissioned you and other leaders in the educational research, psychology and educational testing fields to contribute information, materials and acumen to the collective work entitled *Standards for Educational and Psychological Testing* (the "Standards"). Since its publication in 1999, the Publishers had and still have the right to use the Standards in print, electronic format any and all other formats then known, now known or hereafter to become known. We are confirming in this letter that you accepted this assignment subject to the following terms and conditions:

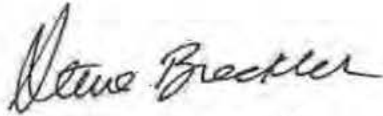
1. You delivered manuscript(s) or review(s) of manuscript(s) within the time period established by Publishers. You were reimbursed for reasonable expenses in connection with your work on the Standards, if approved by the Publishers in advance, upon your submission of receipts. Your name appeared in the Preface of the Standards, showing that you were one of its contributors, and you received a free copy of the Standards upon its publication.
2. You acknowledge that the Standards, and all contributions that you made toward completion and publication of the Standards, was and still is considered a "work made for hire" within the meaning of the United States copyright laws, and that the Publishers own all right, title and interest in and to the copyright in the Standards. To the extent that the Standards were not or are not deemed to be a "work made for hire" you hereby assign nunc pro tunc (now for then) to the Publishers all right, title and interest in and to the Standards.
3. Accordingly, the Publishers may also use the Standards for any and all uses and products, in any and all formats now known or hereafter to become known, including but not limited to print, recorded on hard storage media (e.g., CDs, DVDs, etc.), the Internet and online services.
4. You have granted the Publishers the right to use your name in the Standards, in advertising and promotion related to the Standards, and in any and all ancillary products related to the Standards regardless of the formats in which such use occurs.
5. Your contributions to the Standards were wholly original material not published elsewhere (except for material in the public domain or used with the permission of the owner), did not and does not infringe any copyright, and did not and does not constitute a defamation, or invasion of the right of privacy or publicity, or infringement of any other kind, of any third party.
6. It is specifically understood and intended that you are an independent contractor, and nothing herein is intended or shall be deemed to make you an employee of any of the Publishers.

If the foregoing accurately sets forth our understanding, please sign and date, then scan and return this letter via e-mail attachment to Marianne Ernesto, Director, Testing and Assessment, American Psychological Association, at mernesto@apa.org.

Sincerely,



AMERICAN EDUCATIONAL RESEARCH ASSOCIATION

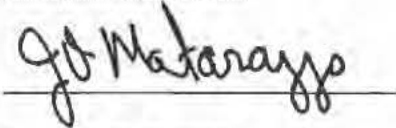


AMERICAN PSYCHOLOGICAL ASSOCIATION



NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

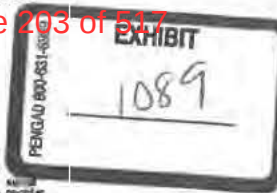
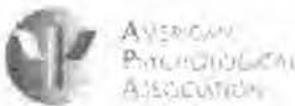
Accepted and Agreed:



Date: April 28, 2014

EXHIBIT EEE

Case No. 1:14-cv-00857-TSC-DAR



April 21, 2014

Dear Member of the Joint Committee to Revised the 1985 *Standards for Educational and Psychological Testing*:

In 1998 and 1999, we, the American Educational Research Association ("AERA"), the American Psychological Association ("APA") and the National Council on Measurement in Education ("NCME"), (collectively, the "Publishers"), commissioned you and other leaders in the educational research, psychology and educational testing fields to contribute information, materials and acumen to the collective work entitled *Standards for Educational and Psychological Testing* (the "Standards"). Since its publication in 1999, the Publishers had and still have the right to use the Standards in print, electronic format any and all other formats then known, now known or hereafter to become known. We are confirming in this letter that you accepted this assignment subject to the following terms and conditions:

1. You delivered manuscript(s) or review(s) of manuscript(s) within the time period established by Publishers. You were reimbursed for reasonable expenses in connection with your work on the Standards, if approved by the Publishers in advance, upon your submission of receipts. Your name appeared in the Preface of the Standards, showing that you were one of its contributors, and you received a free copy of the Standards upon its publication.
2. You acknowledge that the Standards, and all contributions that you made toward completion and publication of the Standards, was and still is considered a "work made for hire" within the meaning of the United States copyright laws, and that the Publishers own all right, title and interest in and to the copyright in the Standards. To the extent that the Standards were not or are not deemed to be a "work made for hire" you hereby assign nunc pro tunc (now for then) to the Publishers all right, title and interest in and to the Standards.
3. Accordingly, the Publishers may also use the Standards for any and all uses and products, in any and all formats now known or hereafter to become known, including but not limited to print, recorded on hard storage media (e.g., CDs, DVDs, etc.), the Internet and online services.
4. You have granted the Publishers the right to use your name in the Standards, in advertising and promotion related to the Standards, and in any and all ancillary products related to the Standards regardless of the formats in which such use occurs.
5. Your contributions to the Standards were wholly original material not published elsewhere (except for material in the public domain or used with the permission of the owner), did not and does not infringe any copyright, and did not and does not constitute a defamation, or invasion of the right of privacy or publicity, or infringement of any other kind, of any third party.
6. It is specifically understood and intended that you are an independent contractor, and nothing herein is intended or shall be deemed to make you an employee of any of the Publishers.

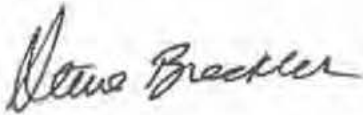
Confidential

If the foregoing accurately sets forth our understanding, please sign and date, then scan and return this letter via e-mail attachment to Marianne Ernesto, Director, Testing and Assessment, American Psychological Association, at mernesto@apa.org.

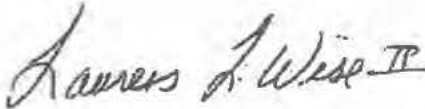
Sincerely,



AMERICAN EDUCATIONAL RESEARCH ASSOCIATION



AMERICAN PSYCHOLOGICAL ASSOCIATION



NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

Accepted and Agreed:



Date: 9/8/14

EXHIBIT FFF

Case No. 1:14-cv-00857-TSC-DAR

EXHIBIT

1090



AMERICAN
PSYCHOLOGICAL
ASSOCIATION



April 21, 2014

Dear Member of the Joint Committee to Revised the 1985 *Standards for Educational and Psychological Testing*:

In 1998 and 1999, we, the American Educational Research Association ("AERA"), the American Psychological Association ("APA") and the National Council on Measurement in Education ("NCME"), (collectively, the "Publishers"), commissioned you and other leaders in the educational research, psychology and educational testing fields to contribute information, materials and acumen to the collective work entitled *Standards for Educational and Psychological Testing* (the "Standards"). Since its publication in 1999, the Publishers had and still have the right to use the Standards in print, electronic format any and all other formats then known, now known or hereafter to become known. We are confirming in this letter that you accepted this assignment subject to the following terms and conditions:

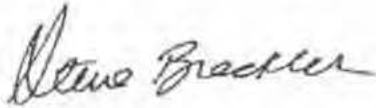
1. You delivered manuscript(s) or review(s) of manuscript(s) within the time period established by Publishers. You were reimbursed for reasonable expenses in connection with your work on the Standards, if approved by the Publishers in advance, upon your submission of receipts. Your name appeared in the Preface of the Standards, showing that you were one of its contributors, and you received a free copy of the Standards upon its publication.
2. You acknowledge that the Standards, and all contributions that you made toward completion and publication of the Standards, was and still is considered a "work made for hire" within the meaning of the United States copyright laws, and that the Publishers own all right, title and interest in and to the copyright in the Standards. To the extent that the Standards were not or are not deemed to be a "work made for hire" you hereby assign nunc pro tunc (now for then) to the Publishers all right, title and interest in and to the Standards.
3. Accordingly, the Publishers may also use the Standards for any and all uses and products, in any and all formats now known or hereafter to become known, including but not limited to print, recorded on hard storage media (e.g., CDs, DVDs, etc.), the Internet and online services.
4. You have granted the Publishers the right to use your name in the Standards, in advertising and promotion related to the Standards, and in any and all ancillary products related to the Standards regardless of the formats in which such use occurs.
5. Your contributions to the Standards were wholly original material not published elsewhere (except for material in the public domain or used with the permission of the owner), did not and does not infringe any copyright, and did not and does not constitute a defamation, or invasion of the right of privacy or publicity, or infringement of any other kind, of any third party.
6. It is specifically understood and intended that you are an independent contractor, and nothing herein is intended or shall be deemed to make you an employee of any of the Publishers.

If the foregoing accurately sets forth our understanding, please sign and date, then scan and return this letter via e-mail attachment to Marianne Ernesto, Director, Testing and Assessment, American Psychological Association, at mernesto@apa.org.

Sincerely,



AMERICAN EDUCATIONAL RESEARCH ASSOCIATION



AMERICAN PSYCHOLOGICAL ASSOCIATION



NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

Accepted and Agreed:

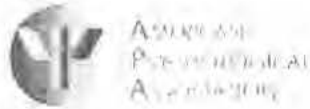


Date: October 21, 2014

EXHIBIT GGG

Case No. 1:14-cv-00857-TSC-DAR

EXHIBIT
1091
PENGAD 800-831-1111



April 21, 2014

Dear Member of the Joint Committee to Revised the 1985 *Standards for Educational and Psychological Testing*:

In 1998 and 1999, we, the American Educational Research Association ("AERA"), the American Psychological Association ("APA") and the National Council on Measurement In Education ("NCME"), (collectively, the "Publishers"), commissioned you and other leaders in the educational research, psychology and educational testing fields to contribute information, materials and acumen to the collective work entitled *Standards for Educational and Psychological Testing* (the "Standards"). Since its publication in 1999, the Publishers had and still have the right to use the Standards in print, electronic format any and all other formats then known, now known or hereafter to become known. We are confirming in this letter that you accepted this assignment subject to the following terms and conditions:

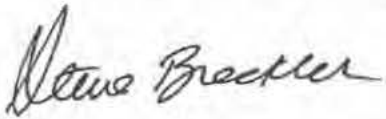
1. You delivered manuscript(s) or review(s) of manuscript(s) within the time period established by Publishers. You were reimbursed for reasonable expenses in connection with your work on the Standards, if approved by the Publishers in advance, upon your submission of receipts. Your name appeared in the Preface of the Standards, showing that you were one of its contributors, and you received a free copy of the Standards upon its publication.
2. You acknowledge that the Standards, and all contributions that you made toward completion and publication of the Standards, was and still is considered a "work made for hire" within the meaning of the United States copyright laws, and that the Publishers own all right, title and interest in and to the copyright in the Standards. To the extent that the Standards were not or are not deemed to be a "work made for hire" you hereby assign nunc pro tunc (now for then) to the Publishers all right, title and interest in and to the Standards.
3. Accordingly, the Publishers may also use the Standards for any and all uses and products, in any and all formats now known or hereafter to become known, including but not limited to print, recorded on hard storage media (e.g., CDs, DVDs, etc.), the Internet and online services.
4. You have granted the Publishers the right to use your name in the Standards, in advertising and promotion related to the Standards, and in any and all ancillary products related to the Standards regardless of the formats in which such use occurs.
5. Your contributions to the Standards were wholly original material not published elsewhere (except for material in the public domain or used with the permission of the owner), did not and does not infringe any copyright, and did not and does not constitute a defamation, or invasion of the right of privacy or publicity, or infringement of any other kind, of any third party.
6. It is specifically understood and intended that you are an independent contractor, and nothing herein is intended or shall be deemed to make you an employee of any of the Publishers.

If the foregoing accurately sets forth our understanding, please sign and date, then scan and return this letter via e-mail attachment to Marianne Ernesto, Director, Testing and Assessment, American Psychological Association, at mernesto@apa.org.

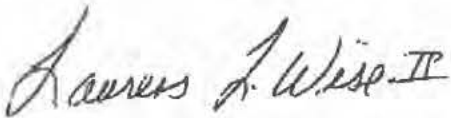
Sincerely,



AMERICAN EDUCATIONAL RESEARCH ASSOCIATION

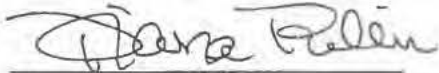


AMERICAN PSYCHOLOGICAL ASSOCIATION



NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

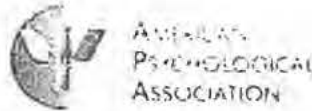
Accepted and Agreed:



Date: 4/23/14

EXHIBIT HHH

Case No. 1:14-cv-00857-TSC-DAR



April 21, 2014

Dear Member of the Joint Committee to Revised the 1985 *Standards for Educational and Psychological Testing*:

In 1998 and 1999, we, the American Educational Research Association ("AERA"), the American Psychological Association ("APA") and the National Council on Measurement in Education ("NCME"), (collectively, the "Publishers"), commissioned you and other leaders in the educational research, psychology and educational testing fields to contribute information, materials and acumen to the collective work entitled *Standards for Educational and Psychological Testing* (the "Standards"). Since its publication in 1999, the Publishers had and still have the right to use the Standards in print, electronic format any and all other formats then known, now known or hereafter to become known. We are confirming in this letter that you accepted this assignment subject to the following terms and conditions:


1. You delivered manuscript(s) or review(s) of manuscript(s) within the time period established by Publishers. You were reimbursed for reasonable expenses in connection with your work on the Standards, if approved by the Publishers in advance, upon your submission of receipts. Your name appeared in the Preface of the Standards, showing that you were one of its contributors, and you received a free copy of the Standards upon its publication.
2. You acknowledge that the Standards, and all contributions that you made toward completion and publication of the Standards, was and still is considered a "work made for hire" within the meaning of the United States copyright laws, and that the Publishers own all right, title and interest in and to the copyright in the Standards. To the extent that the Standards were not or are not deemed to be a "work made for hire" you hereby assign nunc pro tunc (now for then) to the Publishers all right, title and interest in and to the Standards.
3. Accordingly, the Publishers may also use the Standards for any and all uses and products, in any and all formats now known or hereafter to become known, including but not limited to print, recorded on hard storage media (e.g., CDs, DVDs, etc.), the Internet and online services.
4. You have granted the Publishers the right to use your name in the Standards, in advertising and promotion related to the Standards, and in any and all ancillary products related to the Standards regardless of the formats in which such use occurs.
5. Your contributions to the Standards were wholly original material not published elsewhere (except for material in the public domain or used with the permission of the owner), did not and does not infringe any copyright, and did not and does not constitute a defamation, or invasion of the right of privacy or publicity, or infringement of any other kind, of any third party.
6. It is specifically understood and intended that you are an independent contractor, and nothing herein is intended or shall be deemed to make you an employee of any of the Publishers.

If the foregoing accurately sets forth our understanding, please sign and date, then scan and return this letter via e-mail attachment to Marianne Ernesto, Director, Testing and Assessment, American Psychological Association, at mernesto@apa.org.

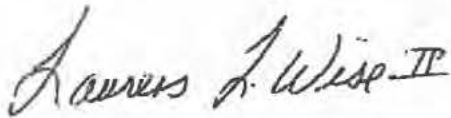
Sincerely,



AMERICAN EDUCATIONAL RESEARCH ASSOCIATION



AMERICAN PSYCHOLOGICAL ASSOCIATION



NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

Accepted and Agreed:



Date: 4/24/14

UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA

AMERICAN EDUCATIONAL RESEARCH) ASSOCIATION, INC., AMERICAN) PSYCHOLOGICAL ASSOCIATION, INC.,) and NATIONAL COUNCIL ON) MEASUREMENT IN EDUCATION, INC.,) Plaintiffs,) v.) PUBLIC.RESOURCE.ORG, INC.,) Defendant.)	Civil Action No. 1:14-cv-00857-TSC-DAR DECLARATION OF LAURESS L. WISE IN SUPPORT OF PLAINTIFFS' MOTION FOR SUMMARY JUDGMENT AND ENTRY OF A PERMANENT INJUNCTION
---	---

I, LAURESS L. WISE, declare:

1. I am the Immediate Past President of the National Council on Measurement in Education, Inc. ("NCME"). I have been a member of this organization for approximately 30 years. I previously was the President of NCME from April 2014 through April 2015, and Vice President of this organization from April 2013 through April 2014. I submit this Declaration in support of the motion of the American Educational Research Association, Inc. ("AERA"), the American Psychological Association, Inc. ("APA"), and the NCME (collectively, "Plaintiffs" or "Sponsoring Organizations") for summary judgment and the entry of a permanent injunction.

2. I also am a principal scientist with the Human Resources Research Organization ("HumRRO"), spending full time on research and evaluation projects relating to educational measurement. I previously served as HumRRO CEO for 13 years, combining management and research activities and, before that, directed research and development for the Armed Services Vocational Aptitude Battery for the Department of Defense. Before that I spent 16 years as a researcher for the American Institutes for Research, rising to the position of Director of Research. I am also a member of both AERA and APA.

3. NCME is a District of Columbia not-for-profit corporation.

4. NCME is a professional organization for individuals involved in assessment, evaluation, testing, and other aspects of educational measurement. NCME's members are involved in the construction and use of standardized tests; new forms of assessment, including performance-based assessment; program design; and program evaluation.

5. In 1955, AERA and NCME prepared and published a companion document to APA's "Technical Recommendations for Psychological Tests and Diagnostic Techniques" (published in 1954), entitled "Technical Recommendations for Achievement Tests."

6. Subsequently, a joint committee of the three organizations modified, revised and consolidated the two documents into the first Joint Standards. Beginning with the 1966 revision, the Sponsoring Organizations collaborated in developing the "Joint Standards" (or simply, the "Standards"). Each subsequent revision of the Standards has been careful to note that it is a revision and update of that document.

7. Beginning in the mid-1950s, the Sponsoring Organizations formed and periodically reconstituted a committee of experts in psychological and educational assessment, charged with the initial development of the Technical Recommendations and then each subsequent revision of the (renamed) Standards. These committees were formed by the three organizations' Presidents (or their designees), who would meet and jointly agree on the membership. Often a chair or co-chairs of these committees were selected by joint agreement. Beginning with the 1966 version of the Standards, this committee became referred to as the "Joint Committee." For example, I was the co-chair of the Joint Committee for the 2014 edition of the Standards.

8. Financial and operational oversight for the Standards' revisions, promotion,

distribution, and for the sale of the 1999 and 2014 Standards has been undertaken by a periodically reconstituted Management Committee, comprised of designees of the three Sponsoring Organizations.

9. All members of the Joint Committee(s) and the Management Committee(s) are unpaid volunteers. The expenses associated with the ongoing development and publication of the Standards include travel and lodging expenses (for the Joint Committee and Management Committee members), support staff time, printing and shipment of bound volumes, and advertising costs.

10. Many different fields of endeavor rely on assessments. The Sponsoring Organizations have ensured that the range of these fields of endeavor is represented in the Joint Committee's membership – *e.g.*, admissions, achievement, clinical counseling, educational, licensing-credentialing, employment, policy, and program evaluation. Similarly, the Joint Committee's members represent expertise across major functional assessment areas – *e.g.*, validity, equating, reliability, test development, scoring, reporting, interpretation, large scale interpolation and cognitive behavioral therapy.

11. From the time of their initial creation to the present, the preparation and periodic revisions to the Standards entail intensive labor and considerable cross-disciplinary expertise. Each time the Standards are revised, the Sponsoring Organizations select and arrange for meetings of the leading authorities in psychological and educational assessments (known as the Joint Committee). During these meetings, certain Standards are combined, pared down, and/or augmented, others are deleted altogether, and some are created as whole new individual Standards. The 1999 version of the Standards is nearly 200 pages, and took more than five years to complete – resulting from work put in by the Joint Committee to generate a set of best

practices on educational and psychological testing that are respected and relied upon by leaders in their fields.

12. The Standards originally were created as principles and guidelines – a set of best practices to improve professional practice in testing and assessment across multiple settings, including education and various areas of psychology. The Standards can and should be used as a recommended course of action in the sound and ethical development and use of tests, and also to evaluate the quality of tests and testing practices. Additionally, an essential component of responsible professional practice is maintaining technical competence. Many professional associations also have developed standards and principles of technical practice in assessment. The Sponsoring Organizations' Standards have been and still are used for this purpose.

13. The Standards, however, are not simply intended for members of the Sponsoring Organizations, AERA, APA, and NCME. The intended audience of the Standards is broad and cuts across audiences with varying backgrounds and different training. For example, the Standards also are intended to guide test developers, sponsors, publishers, and users by providing criteria for the evaluation of tests, testing practices, and the effects of test use. Test user standards refer to those standards that help test users decide how to choose certain tests, interpret scores, or make decisions based on tests results. Test users include clinical or industrial psychologists, research directors, school psychologists, counselors, employment supervisors, teachers, and various administrators who select or interpret tests for their organizations. There is no mechanism, however, to enforce compliance with the Standards on the part of the test developer or test user. The Standards, moreover, do not attempt to provide psychometric answers to policy or legal questions.

14. The Standards promote the development of high quality tests and the sound use of

results from such tests. Without such high quality standards, tests might produce scores that are not defensible or accurate, not an adequate reflection of the characteristic they were intended to measure, and not fair to the person tested. Consequently, decisions about individuals made with such test scores would be no better, or even worse, than those made with no test score information at all. Thus, the Standards help to ensure that measures of student achievement are relevant, that admissions decisions are fair, that employment hiring and professional credentialing result in qualified individuals being selected, and patients with psychological needs are diagnosed properly and treated accordingly. Quality tests protect the public from harmful decision making and provide opportunities for education and employment that are fair to all who seek them.

15. The Standards apply broadly to a wide range of standardized instruments and procedures that sample an individual's behavior, including tests, assessments, inventories, scales, and other testing vehicles. The Standards apply equally to standardized multiple-choice tests, performance assessments (including tests comprised of only open-ended essays), and hands-on assessments or simulations. The main exceptions are that the Standards do not apply to unstandardized questionnaires (*e.g.*, unstructured behavioral checklists or observational forms), teacher-made tests, and subjective decision processes (*e.g.*, a teacher's evaluation of students' classroom participation over the course of a semester).

16. The Standards have been used to develop testing guidelines for such activities as college admissions, personnel selection, test translations, test user qualifications, and computer-based testing. The Standards also have been widely cited to address technical, professional, and operational norms for all forms of assessments that are professionally developed and used in a variety of settings. The Standards additionally provide a valuable public service to state and

federal governments as they voluntarily choose to use them. For instance, each testing company, when submitting proposals for testing administration, instead of relying on a patchwork of local, or even individual and proprietary, testing design and implementation criteria, may rely instead on the Sponsoring Organizations' Standards to afford the best guidance for testing and assessment practices.

17. The Standards were not created or updated to serve as a legally binding document, in response to an expressed governmental or regulatory need, nor in response to any legislative action or judicial decision. However, the Standards have been cited in judicial decisions related to the proper use and evidence for assessment, as well as by state and federal legislators. These citations in judicial decisions and during legislative deliberations occurred without any lobbying by the Plaintiffs.

18. NCME has never communicated with Congress for the purpose of encouraging the enactment of the Standards into law.

19. Additionally, NCME has never solicited any government agency to incorporate the Standards into the Code of Federal Regulations or other rules of Federal or State agencies.

20. In the policymaking arena, NCME believes the Standards should be treated as guidelines informing the enactment of legislation and regulations consistent with best practices in the development and use of tests – to insure that they are valid, reliable and fair.

21. The Sponsoring Organizations promote and sell copies of the Standards via referrals to the AERA website, at annual meetings, in public offerings to students, and to educational institution faculty. Advertisements promoting the Standards have appeared in meeting brochures, in scholarly journals, and in the hallways at professional meetings. Accompanying this Declaration as Exhibit KKK is a true copy of advertisements for the 1999

Standards published in NCME's Journal of Educational Management. These advertisements were produced at Bates Nos. AERA_APA_NCME_0031444-0031451.

22. Distribution of the Standards is closely monitored by the Sponsoring Organizations. AERA, the designated publisher of the Standards, sometimes does provide promotional complementary print copies to students or professors. Except for these few complementary print copies, however, the Standards are not given away for free; and certainly they are not made available to the public by any of the three organizations for anyone to copy free of charge.

23. To date, NCME has never posted, or authorized the posting of, a digitized copy of the 1999 Standards on any publicly accessible website.

24. Without receiving at least some moderate income from the sales of the Standards to offset their production costs and to allow for further revisions, it is very likely that the Sponsoring Organizations would no longer undertake to periodically update them, and it is unknown who else would.

25. In late 2013 and early 2014, the Sponsoring Organizations became aware that the 1999 Standards had been posted on the Internet without their authorization, and that students were obtaining free copies from the posting source. Upon further investigation, the Sponsoring Organizations discovered that Public Resource was the source of the online posting.

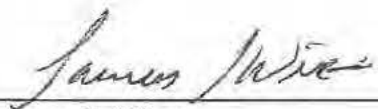
26. Public Resource posted Plaintiffs' 1999 Standards to its website and the Internet Archive website without the permission or authorization of any of the Sponsoring Organizations.

27. Plaintiffs have been made aware that at least some of those users who obtained the 1999 Standards for free from Public Resource did so to avoid paying the modest sale price for authorized print copies.

28. Accompanying this Declaration as Exhibit LLL is a true copy of an e-mail dated March 5, 2014 from Gregory J. Cizek to me regarding a student not purchasing the 1999 Standards because “they [were] available for free online” at <https://law.resource.org/pub/us/cfr/ibr/001/aera.standards.1999.pdf>.” This e-mail exchange was marked as Exhibit 1252 during my deposition.

I DECLARE, under the penalty of perjury, that the foregoing is true and correct.

Dated: December 9, 2015



Laress L. Wise

EXHIBIT 1

LAURESS L. WISE

Curriculum Vitae

OVERVIEW

Dr. Lauress Wise has over 35 years' experience in educational research and continues extensive work on educational policy and assessment issues. Dr. Wise currently advises several states and the PARCC assessment consortium on technical issues in test development and use. He serves on the Board of the National Council of Measurement in Education as the immediate past-president. He is also serving on a National Research Council Committee that is evaluating the NAEP achievement levels. He recently co-chaired the panel that revise the AERA/APA/NCME *Standards for Educational and Psychological Testing* and previously chaired the National Academy of Sciences Board on Testing and Assessment. Recent research and development efforts include a 15-year independent evaluation of the California High School Exit Exam and quality assurance work for the National Assessment of Educational Progress (NAEP). Dr. Wise previously served on several National Research Council committees, chairing the Committees on Scientific Research in Education and the Evaluation of the National Voluntary Tests.

EDUCATION

Ph.D. , Mathematical Psychology	1975	University of California, Berkeley
B.S. , Mathematics, Psychology (with Distinction)	1967	Stanford University

AREAS OF EXPERTISE

- Test Development and Validation
- Program and Policy Evaluation
- Test Use Policy
- Project Management
- Statistical and Psychometric Issues
- Computer-Based Testing

PROFESSIONAL EXPERIENCE

Human Resources Research Organization **1994 - 2015**
Principal Scientist

- Served as HumRRO's president from 1994 to 2007. Remained active in research on testing and test use policy. Directed two major HumRRO educational testing projects, one to provide quality assurance for the National Assessment of Educational Progress (NAEP) and the other an independent evaluation of California's High School Exit Exam (CAHSEE). He continues to serve as a senior psychometric advisor for a graduate school admissions testing program.
- Served as co-chair of the committee that revised the 1999 AERA/APA/NCME Standards for Educational and Psychological Testing, and has previously served as Chair of the National Academy of Science (NAS) Board on Testing and Assessment and chaired the NAS Committee on Research in Education.
- Currently serves on technical advisory committees for the Hawaii, Wyoming, Utah, Tennessee, and Virginia departments of education, and the Partnership for Assessing

Readiness for College and Career (PARCC) advisory committees. Also serves on the Rhode Island Technical Advisory Committee for Teacher Evaluation.

- Served as co-Principal Investigator on the first year of the Congressionally-mandated evaluation of President Clinton's Voluntary National Tests and chaired the NAS committee that performed the second year of that evaluation, and on the NAS committee to evaluate the NAEP and on the National Academy of Education's Panel for the Evaluation of the NAEP Trial State NAEP.
- Other work includes vertical alignment of state content standards, modeling the effects of motivation on examinee performance on low-stakes assessments, the impact of changes in exclusions on NAEP results for Kentucky, and scaling constructed response and multiple choice items on the Florida assessment. Dr. Wise also worked on the development and validation of a computer-administered assessment now used for selection of air traffic controllers and developed a computer-based system for assessing work values as part of a Department of Labor effort to develop improved career guidance tools.

Defense Manpower Data Center

1990 - 1994

Chief, Personnel Testing Division

- Spokesperson for the Department of Defense on matters relating to the development and use of cognitive tests. Dr. Wise's unit was responsible for all research and development for the Armed Services Vocational Aptitude Battery (ASVAB).
- Work included evaluation and implementation of a computerized, adaptive version of the ASVAB, automated item and form development procedures, new career exploration procedures for use with the high school testing program, the development and testing of a new career interest inventory, and extensive validity research.

American Institutes for Research (AIR)

1974 - 1990

Associate Research Scientist to Director of Research

- Directed a variety of studies and projects, including the Review and Analysis of the General Aptitude Test Battery (GATB) Project for the U.S. Employment Service within the Department of Labor, the Army Synthetic Validation Project, the Army's Computerized Adaptive Screening Test (CAST) Revision Project, and validation studies for the Software Engineering Institute at Carnegie Mellon University.
- Served as director of analysis for the Army's massive Project A, analyzing new selection tests, developing models of performance in a variety of jobs, and assessing the validity of each new test for predicting different facets of performance in different jobs.
- Served for twelve years as the chief psychometrician for the Medical College Admissions Test, developing procedures for the screening and calibration of new items and for the construction and equating of new forms. Also consulted with the Department of Education on issues related to testing and data analysis as part of the Statistical Analysis Group in Education.
- From 1978 to 1982, served as Director of Project TALENT, a nationally representative longitudinal study of nearly 400,000 members of the high school classes of 1960 through 1963. Oversaw the collection of the final wave of follow-up data and conducted targeted research on issues such as gender differences in mathematics achievement, school differences in student achievement, the development of careers in science and medicine, and the consequences of adolescent childbearing.

University of California

Programmer and Instructor

- While a graduate student at the University of California, Dr. Wise served as the computer consultant for the Psychology Department, helping both faculty and students in the design and execution of data analyses. He also taught an undergraduate course in Psychological Statistics.

California Department of Public Health

1968 - 1972

Computer Programmer and Data Processing Systems Analyst

- Created data systems to support licensing functions and vital statistics systems at the Department of Public Health. Was in-house project manager for a new management information system that involved defining "outputs" for each bureau and department and relating these outputs to costs.

PROFESSIONAL AFFILIATIONS AND SERVICE

- American Educational Research Association (AERA)
- American Psychological Association (APA)
 - Divisions 5, 14, 19
- Member, National Council on Measurement in Education (NCME)
- Psychometric Society

SELECTED BIBLIOGRAPHY

Publications

Wise, L. L. (in press). How we got to where we are: Evolving policy demands for the next generation assessments. In *Next Generation Assessments*. R. Lissitz, Ed. (To be published in 2016).

Wise, L. L. & Plake, B. S. (2015). "Test design and development following the *Standards for Educational and Psychological Testing*". In Lane, S., Haladyna, T., and Raymond, M. (Eds.). *Handbook of Test Development*. New York, NY: Routledge.

Plake, B. S., & Wise, L. L. (2014). Revision of the AERA, APA, NCME *Standards for Educational and Psychological Testing*: What is their role and importance for NCME. *Educational Measurement: Issues and Practice*, 33(4), 4-12.

Wise, L. L. (2010). Accessible Reading Assessments for Students with Disabilities: Summary and Conclusions. *Applied Measurement in Education*, 23(2), 209-214.

Wise, L. L. (2006). Encouraging and Supporting Compliance with Standards for Educational Tests. *Educational Measurement: Issues and Practice* 25(3), 51-53.

National Research Council. (2005). *Advancing Scientific Research in Education*. Committee on Research in Education. Lisa Towne, Laress L. Wise, and Tina M Winders, Editors. Center for

Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academy Press.

National Research Council. (2004). *Strengthening Peer Review in Federal Agencies That Support Education Research*. Committee on Research in Education. L. Towne, J.M. Fletcher, and L.L. Wise, Eds. Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Wise, L.L. (2004). The National Assessment of Educational Progress - what it tells educators. In J.E. Wall & G.R. Walz (Eds.). *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 729-741). Greensboro, NC: CAPS Press.

Wise, L.L. & Hoffman, R.G. (2002). *How will assessment data to be used to document the impact of educational reform*. In R. W. Lissitz and W. D. Schafer (Eds.) *Assessment in Educational Reform: Both means and ends*. Boston, MA: Allyn & Bacon.

Wise, L.L., Noeth, R.J., & Koenig, J.A. (Eds.) (1999). *Evaluation of the voluntary national tests, year 2 interim report*. National Research Council. Washington, DC: National Academy Press.

Wise, L.L., Hauser, R.M., Mitchell, K.J., & Feuer, M.J. (1999). *Evaluation of the voluntary national tests: Phase I*. National Research Council, Commission on Behavioral and Social Sciences and Education, Board on Testing and Assessment. Washington, DC: National Academy Press.

Wise, L.L., Curran, L.T. Curran, & McBride, J.R. (1997). *CAT-ASVAB cost and benefit analyses*. In W.A. Sands, B.K. Waters, & J.R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation*. Washington, DC: American Psychological Association.

Wolfe, J.H., Alderton, D.L., Larson, G.E., Bloxom, B., & Wise, L.L. (1997). *Expanding the content of CAT-ASVAB: New tests and their validity*. In W.A. Sands, B.K. Waters, & J.R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation*. Washington, DC: American Psychological Association.

Wall, J.E., Wise, L.L., & Baker, H.E. (1996). *Development of the interest-finder: A new RIASEC-based interest inventory*. *Measurement and Evaluation in Counseling and Development*, 29, 134-152.

Wise, L.L. (1994). *Goals of the selection and classification decision*. In M.G. Rumsey, C.B. Walker, & J.H. Harris (Eds.). *Personnel selection and classification*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wise, L.L. (1994). *Setting performance goals for the DOD Linkage Model*. In B.F. Green & A.S. Mavor (Eds.) *Modeling cost and performance for military enlistment*. Washington, DC: National Academy of Sciences Press.

Rudner, L.M., Wise, L.L., & Stonehill, R.M. (1991). The ERIC Clearinghouse on Tests, Measurement, and Evaluation (ERIC/TME) -- A growing resource. *Applied Measurement in Education*, 4, 1-10.

Wise, L.L. (1991). *The validity of test scores for selecting and classifying enlisted recruits*. In B.R. Gifford & L.C. Wing (Eds.), *Test policy in Defense: Lessons from the military for education, training, and employment*. Boston, MA: Kluwer Academic Press.

Campbell, J.P., McHenry, J.J., & Wise, L.L. (1990). *Modeling job performance in a population of jobs*. *Personnel Psychology*, 43, 313-334.

Young, W.Y., Houston, J.S., Harris, J.H., Hoffman, R.G., & Wise, L.L. (1990). *Large-scale predictor validation in Project A: Data collection procedures and data base preparation*. *Personnel Psychology*, 43, 301-312.

Wise, L.L., McHenry, J.J., & Campbell, J.P. (1990). *Identifying optimal predictor composites and testing for generalizability across jobs and performance factors*. *Personnel Psychology*, 43, 355-366.

Wise, L.L., Campbell, J.P., & Arabian, J.M. (1988). *The Army Synthetic Validation Project*. In B.F. Green, H. Wing, & A.K. Wigdor (Eds.) *Linking military enlistment standards to job performance*. Washington, DC: National Academy Press.

Wise, L.L. (1985). *Project TALENT: Mathematics course participation in the 1960's and its career consequences*. In S.F. Chipman, L.R. Brush, & D.M. Wilson (Eds.), *Women and mathematics: Balancing the equation*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Abeles, R.P., Steel, L.M., & Wise, L.L. (1980). *Patterns and implications of life-course organization: Project TALENT studies*. In P.B. Baltes & O.G. Brim, Jr. (Eds.), *Lifespan development and behavior* (Vol. III). New York: Academic Press.

Wise, L.L., & Steel, L.M. (1980). *Educational attainment of the high school classes of 1960 through 1963: Findings from Project TALENT*. In A.C. Kerckhoff (Ed.), *Longitudinal perspectives on educational attainment*. Greenwich, CT: JAI Press.

Wise, L.L. (1979). *Project TALENT: Studying the development of our human resource*. In J.E. Milholland (Ed.), *New directions for testing and measurement: Insights from large-scale surveys*. San Francisco: Jossey-Bass Inc.

Thacker, A. A., Dickinson, E. R., Bynum, B. H., Wen, Y., Smith, E. A., Sinclair, A. L., Deatz, R. C., & Wise, L. L. (2015). *Findings from the quality of items/tasks/stimuli investigations: PARCC field tests* (2015 No. 009). Alexandria, VA: Human Resources Research Organization.

Becker, D.E., Wise, L. L., Hardoin, M. M. & Watters, C. (2014). *Independent evaluation of the California High School Exit Examination: 2014 Biennial report* (2014 No. 001). Alexandria, VA: Human Resources Research Organization.

Thacker, A. A., Dickinson, E. R., Wise, L. L., & Becker, D. E. (2014). *PARCC studies to examine comparability of scores across states, assessment forms, scoring methods and other relevant variables memorandum* (2014 No. 005). Alexandria, VA: Human Resources Research Organization.

Thacker, A. A., Sinclair, A. L., Wise, L. L., & Becker, D. E., (2014). *PARCC validity studies including predictive and longitudinal studies memorandum* (2014 No. 020). Alexandria, VA: Human Resources Research Organization.

Becker, D. E., Wise, L. L., Hardoin, M. M., & Watters, C. (Eds.) (2013). *Independent evaluation of the California High School Exit Examination: 2013 evaluation report* (2013 No. 062). Alexandria, VA: Human Resources Research Organization.

Thacker, A. A., & Wise, L. L. (2013). *Independent review of New York item quality and item screening processes: Summary of findings for New York State Department of Education (NYSD)* (2013 No. 015). Alexandria, VA: Human Resources Research Organization.

Wise, L. L., Becker, D. E., Diaz, T. E., Buckland, W. W., & Norman, R. L. (2013). *Quality assurance for NAEP 2013 reading results* (2013 No. 049). Alexandria, VA: Human Resources Research Organization.

Becker, D. E., Wise, L. L., Hardoin, M. M., & Waters, C. (Eds.). (2012). *Independent evaluation of the California High School Exit Examination: 2012 biennial report* (FR-11-82). Alexandria, VA: Human Resources Research Organization. [Sixth in a series of biennial reports]

Becker, D. E., Wise, L. L., Hardoin, M. M., & Watters, C. (Eds.). (2012). *Independent evaluation of the California High School Exit Examination: 2011 evaluation report* (FR-12-54). Alexandria, VA: Human Resources Research Organization. [Thirteenth in a series of annual reports]

Ramsberger, P. F., Knapp, D. J., & Wise, L. L. (2012). *Overview of procedures used in high-stakes testing programs* (FR-12-07). Alexandria, VA: Human Resources Research Organization.

Thacker, A. A., Dickinson, E. R., Sinclair, A. L., & Wise, L. L. (2012). *Independent review of test item quality for New York State Department of Education (NYSED)* (FR-12-42). Alexandria, VA: Human Resources Research Organization.

Wise, L. L., Thacker, A. A., & Hoffman, R. G. (2012). *Independent review of 2012 equating process for the New York State Department of Education (NYSED)* (FR-12-37). Alexandria, VA: Human Resources Research Organization.

Becker, D. E., Wise, L. L., & Sellman, W. S. (2011). *Comprehensive review of NAEP quality control procedures and documentation: Fiscal Year 2011* (FR-11-17). Alexandria, VA: Human Resources Research Organization.

Trippe, D. M., Waugh, G. W., Hoffman, R. G., McCloy, R. A., & Wise, L. L. (2010). *Computer-based testing (CBT) feasibility study for the CFP[®] certification examination* (FR-10-78). Alexandria, VA: Human Resources Research Organization.

Becker, D.E., Wise, L.L., Hardoin, M.M., & Watters, C. (Eds.) (2012). *Independent evaluation of the California high school exit examination: 2011 evaluation report* (FR-12-54). Alexandria, VA: Human Resources Research Organization.

Thacker, A.A., Dickinson, E.R., Sinclair, A.L., & Wise, L.L. (2012). *Independent review of test item quality for New York State Department of Education (NYSED)* (FR-12-42). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Thacker, A.A., Hoffman, R.G. (2012). *Independent review of 2012 equating process* (FR-12-37). Alexandria, VA: Human Resources Research Organization.

Becker, D.E., Wise, L.L., Hardoin, M.M., & Waters, C. (Eds.) (2012). *Independent evaluation of the California High School Exit Examination: 2012 biennial report* (FR-11-82). Alexandria, VA: Human Resources Research Organization.

Becker, D.E., Wise, L.L., Hardoin, M.M., & Waters, C. (Eds.) (2011). *Independent evaluation of the California High School Exit Examination: 2011 evaluation report* (FR-11-51). Alexandria, VA: Human Resources Research Organization.

Becker, D.E., Wise, L.L., & Sellman, W. S. (2011). *Comprehensive review of NAEP quality control procedures and documentation: Fiscal Year 2011* (FR-11-17). Alexandria, VA: Human Resources Research Organization.

Trippe, D.M., Waugh, G.W., Hoffman, R.G., McCloy, R.A., & Wise, L.L. (2010). *Computer-based testing (CBT) feasibility study for the CFP[®] certification examination* (FR-10-78). Alexandria, VA: Human Resources Research Organization.

Becker, D.E., Wise, L.L., & Waters, C. (Eds.) (2010). *Independent evaluation of the CAHSEE 2010 evaluation report* (FR-10-56). Alexandria, VA: Human Resources Research Organization.

Hoffman, R.G., Wise, L.L., Diaz, T.E., Gribben, M.A., & Fry, S.C. (2009). *Verification of NAEP grade 4 mathematics 2007 to 2009 trend results: Report of HumRRO activities* (FR-09-66). Alexandria, VA: Human Resources Research Organization.

Becker, D.E., Wise, L.L., & Watters, C. (Eds.) (2009). *Independent evaluation of the California High School Exit Examination (CAHSEE): 2009 evaluation report* (FR-09-65). Alexandria, VA: Human Resources Research Organization.

Gribben, M., Diaz, T., Wise, L.L. (2009). *TUDA charter school study draft report: Base year special study* (FR-09-13). Alexandria, VA: Human Resources Research Organization.

Becker, D.E., Wise, L.L., & Watters, C. (Eds.) (2008). *Independent evaluation of the California High School Exit Examination (CAHSEE): 2008 evaluation report* (FR 08-100). Alexandria, VA: Human Resources Research Organization

Gribben, M.A., Wise, L.L., & Becker, D.E. (2008). *Review of web-based technical documentation process FY07 NAEP-QA special study report* (TR-08-17). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Becker, D.E., & Campbell, H. (2008). *NAEP-QA review of procedures for mapping state performance standards onto the NAEP scale* (TR-08-05). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., & Hoffman, R.G. (2007). *HumRRO investigation of 2007 NAEP reading gains for American Indian/Alaskan native students* (FR-07-85). Alexandria, VA: Human Resources Research Organization.

Campbell, H.L., Wise, L.L., Becker, D.E., Campbell, C.H. (2007). *NAEP-QA FY07 specialty study: Effects of expert scored papers (ESP) on scoring in the National Assessment of Educational Progress* (FR-07-65). Alexandria, VA: Human Resources Research Organization.

Wise, L.L. (2007). *Review of 2005 and 2006 Terra Nova results for the school district of Philadelphia* (FR-07-18). Alexandria, VA: Human Resources Research Organization.

Wise, L. L., Taylor, L. R., Becker, D. E., Gladden, F. B., Handy, K., Thacker, A. A., Schultz, S., Willison, S., & Dean, J. (2007). *Development of performance level descriptors for the California Standards Tests (CSTs) and the High School Exit Examination (CAHSEE)*. (TR-07-01). Alexandria, VA: Human Resources Research Organization.

Taylor, L. R., Wise, L. L., Thacker, A. A., Moody, R. L., Koger, L. E., Dickinson, E. R., & Trippe, D. M. (2007). *Independent evaluation of the California Standards Tests (CSTs) and the California Alternate Performance Assessment (CAPA)*. (TR-07-04). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Becker, D.E., Butler, F.L., Schantz, L., Bao, H., Sun, S., & Campbell, H.L. (2006). *Independent evaluation of the California High School Exit Examination (CAHSEE): 2006 evaluation report* (FR-06-91). Alexandria, VA: Human Resources Research Organization.

Stawarski, C., Teasdel, T., Wise, L.L., Schultz, S., Maloutas, M. (2006). *National Assessment of Educational Progress (NAEP) State Coordinator Program Evaluation* (IR-06-86). Alexandria, VA: Human Resources Research Organization.

Becker, S., Thacker, A., Campbell, H., & Wise, L.L. (2006). *Answering the billion dollar question: Are students learning more since the implementation of No Child Left Behind? Phase 1 report-draft* (DFR 06-73). Alexandria, VA: Human Resources Research Organization.

Diaz, T., Le, H., & Wise, LL (2006). *NAEP-QA FY05 special study: 12th grade math trend estimates summary draft* (FR-06-43). Alexandria, VA: Human Resources Research Organization.

Sellman, W.S., Becker, D.E., Wise, L.L., & Hoffman, R.G. (2006). *Notes from the meeting of the NAEP-Quality Assurance Consultant Panel, January 25-26, 2006* (SR-06-09). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Le, H., Hoffman, R.G., & Becker, D.E. (2006). *Testing NAEP full population estimates for sensitivity to violation of assumptions: Phase II* (TR-06-08). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Becker, D.E., Harris, C.D., Taylor, L.R., Johnstone, C.J., Miller, N.A., Thompson, S.J., Sun, S., Shen, X., Wang, X., Koger, L.E., & Moody, R. (2006). *Independent evaluation of the California High School Exit Examination (CAHSEE): Third biennial report* (FR-06-02). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Taylor, L.R., Wang, X., Becker, D.E., & Thacker, A. A. (2006). *Review of the appropriateness of the California high school exit exam content standards for high school accountability* (TR-06-01). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., & Becker, D.E. (2005). *Independent Checks of 2005 NAEP State and TUDA Results: Reading and Math in Grades 4 and 8* (FR-05-67). Alexandria, VA: Human Resources Research Organization.

Becker, D.E., Diaz, T.E., Le, H., Shen, X., Hoffman, G.R., & Wise, L.W. (2005). *Independent Checks of 2005 NAEP State and TUDA Results: Draft Report* (FR-05-61). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Becker, D.E., Harris, C.D., Taylor, L.R., Johnstone, C.J., Miller, N.A., Thompson, S.J., Sun, S., Shen, X., Butler, F.L., Wang, X., Koger, L.E., Moody, R., Deatz, R., Koger, M., Dickinson, E., Gensberg, S., Hilton, R.A., Kelley, N.L., & Stevens, C. (2005). *Independent Evaluation of the California High School Exit Examination (CAHSEE) 2005 Evaluation Report* (FR-05-43). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Zhang, L., Winter, P., Taylor, L., & Becker, D.E. (2005). *Vertical Alignment of Grade-Level Expectations for Student Achievement: Report of a Pilot Study* (TR-05-28). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Becker, D.E., Harris, C.D., Sun, S., Wang, X., & Brown, D.G. (2004). *Independent Evaluation of the California High School Exit Examination (CAHSEE): Year 5 Evaluation Report* (FR-04-53). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Hoffman, R.G., & Becker, D.E. (2004). *Testing NAEP full population estimates for sensitivity to violation of assumptions*. Final Report (Draft). (TR-04-27). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Harris, C.D., Koger, L.E., Bacci, E.D., Ford, J.P., Brown, D.G., Becker, D.E., Sun, S., Koger, M.E., Deatz, R.C., & Coumbe, K.L. (2004). *Independent evaluation of the California High School Exit Examination (CAHSEE): Second biennial report* (FR-04-01). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Becker, D.E., & Gribben, M. (2003). *Notes from the meeting of the National Assessment Governing Board, November 13-15, 2003* (SR-03-101). Alexandria, VA: Human Resources Research Organization.

Wise, L. (2003). *Notes from the meeting of the NAEP validity studies panel - September 12, 2003* (SR-03-71). Alexandria, VA: Human Resources Research Organization.

Hoffman, R.G., & Wise, L.L. (2003). *NAEP quality assurance checks for the 2003 reading assessment results for grade 4* (FR-03-66). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Harris, C.D., Brown, D.G., Becker, D.E., Sun, S., & Coumbe, K.L. (2003). *California High School Exit Examination (CAHSEE): Year 4 evaluation report* (FR-03-64r). Alexandria, VA: Human Resources Research Organization.

Wise, L.L. (2003). *Testing NAEP full population estimates for sensitivity to violation of assumptions – draft design and management plan* (FR-03-61). Alexandria, VA: Human Resources Research Organization.

Hoffman, R.G., & Wise, L.L. (2003). *NAEP quality assurance special studies III-B and III-C: 2003 data replication – final design and management plan* (FR-03-59). Alexandria, VA: Human Resources Research Organization.

Hoffman, R.G., Wise, L.L., & Becker, D.E. (2003). *NAEP quality assurance special study III-A: 2002 data replication – final design and management plan* (FR-03-58). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Sellman, S., & Sipes, S. (2003). *Notes from the Meeting of the National Assessment Governing Board May 15-17, 2003* (SR-03-32). Alexandria, VA: Human Resources Research Organization.

Hoffman, R.G., Becker, D.E., & Wise, L.L. (2003). *NAEP quality assurance checks of the 2002 reading assessment results for Delaware* (FR-03-25). Alexandria, VA: Human Resources Research Organization.

Hoffman, R.G., Wise, L.L., & Sticha, P. J. (2003). *Review of NAEP quality control plans* (TR-03-07). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G. & Wise, L. L. (2003). *The accuracy of school classifications for the 2002 accountability cycle of the Kentucky Commonwealth Accountability Testing System* (FR-03-06). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Becker, D.E., & Ramsberger, P.F. (2003). *Report on past NAEP problems* (FR-03-03). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Sipes, D.E., Harris, C.D., Ford, J.P., Sun, S., Dunn, J., & Goldberg, G.L. (2002). *Independent evaluation of the California High School Exit Examination (CAHSEE): Year 3 evaluation report* (IR-02-28). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Sipes, D.E., Harris, C.D., George, C.E., Ford, J.P., & Sun, S. (2002). *Independent evaluation of the California High School Exit Examination (CAHSEE): Analysis of the 2001 administration* (FR-02-02). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Sipes, D.E. (Sunny), George, C.E., Ford, J. P., & Harris, C.D. (2001). *California High School Exit Examination (CAHSEE): Year 2 evaluation report* (IR-01-29). Alexandria, VA: Human Resources Research Organization.

Hoffman, R.G., & Wise, L.L. (2001). *The accuracy of school classifications for the interim accountability cycle of the Kentucky Commonwealth Accountability and Testing System*. (FR-01-26). Alexandria, VA: Human Resources Research Organization.

McBride, J.R., Paddock, A.F., Wise, L.L., Strickland, W.J., & Waters, B.K. (2001). *Testing via the internet: A literature review and analysis of issues for Department of Defense internet testing of the Armed Services Vocational Aptitude Battery (ASVAB) in the high schools* (FR-01-12). Alexandria, VA: Human Resources Research Organization.

Hoffman, R.G., & Wise, L.L. (2000). *The accuracy of students' novice, apprentice, proficient, and distinguished classifications of the Kentucky Core Content Test* (FR-00-25). Alexandria, VA: Human Resources Research Organization.

Hoffman, R.G., & Wise, L.L. (2000). *School classification accuracy final analysis plan for the commonwealth accountability and testing system* (FR-00-26). Alexandria, VA: Human Resources Research Organization.

Hoffman, R.G., Thacker, A., & Wise, L.L. (2000). *The accuracy of students' novice, apprentice, proficient, and distinguished classifications for the 2000 Kentucky Core Content Test* (FR-00-41). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Harris, C.D., Sipes, D.E., Hoffman, R.G., & Ford, J.P. (2000). *High school exit examination (HSEE): Year 1 evaluation report* (IR-00-27r). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Harris, C.D., Sipes, D.E., Collins, M.M., Hoffman, R.G., & Ford, J.P. (2000). *High school exit examination (HSEE): Supplemental year 1 evaluation report* (IR-00-37). Alexandria, VA: Human Resources Research Organization.

Hoffman, R.G., & Wise, L.L. (1999). *Establishing the reliability of student level classifications: Analytic plan and demonstration* (FR-WATSD-99-34). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., Hoffman, R.G., & Thacker, A.A. (August 1999). *Evaluation of calibration and equating procedures for the Florida State Assessment*. (FR-WATSD-99-41). Alexandria, VA: Human Resources Research Organization.

Wise, L.L., McCloy, R.A., & Quartetti, D.A. (1998). *Indicators of student effort on the National Assessment of Educational Progress* (DFR-EADD-98-58). Alexandria, VA: Human Resources Research Organization.

McCloy, R.A., Russell, T.L., & Wise L.L. (Eds.). (1997). *General Aptitude Test Battery (GATB) improvement project final report*. Washington, DC: U.S. Department of Labor, Divisions of Skills Assessment and Analysis, Office of Policy and Research, Employment and Training Administration.

Wise, L.L. (1997, June). *Merging ASVAB and KIRIS on-demand scores: Report of preliminary results* (LRS 97-4). Frankfort, KY: Kentucky Department of Education.

Wise, L.L., Welsh, J., Grafton, F., Foley, P., Earles, J., Sawin, L., & Divgi, D.R. (1992). *Sensitivity and fairness of the Armed Services Vocational Aptitude Battery (ASVAB) technical composites*. Monterey, CA: Defense Manpower Data Center.

Wise, L.L., Chia, W.J., & Rudner, L.M. (1990). *Identifying necessary job skills: A review of previous approaches*. Washington, DC: Pelavin Associates, Inc.

Wise, L.L., Peterson, N.G., Hoffman, R.G., Campbell, J.P., & Arabian, J.M. (1990). *Army Synthetic Validity Project: Report of phase III results*. Washington, DC: American Institutes for Research.

Wise, L.L., McHenry, J.J., Chia, W.J., Szenas, P.L., & McBride, J.R. (1989). *Refinement of the Computer Adaptive Screening Test (CAST)*. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Wise, L.L., Hough, L.M., Szenas, P.L., & Keyes, M.A. (1988). *Phase I report: armed services applicant profile (ASAP) item fairness analysis*. Washington, DC: American Institutes for Research.

Wise, L.L., McHenry, J.J., & Young, W.Y. (1986). *Project A concurrent validation: treatment of missing data* (RS-WP-86-08). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

McLaughlin, D.H., Rossmessl, P.G., Wise, L.L., Brandt, D.A., & Wang, M. (1984). *Validation of current and alternative ASVAB area composites based on training and SQT information on FY 1982 enlisted accessions* (Technical Report No. 651). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Card, J.J., & Wise, L.L. (1981). *Teenage mothers and teenage fathers: The impact of early childbearing on the parents' personal and professional lives*. In F.F. Furstenberg, R. Lincoln, & J. Menken (Eds.), *Teenage sexuality, pregnancy, and childbearing*. Philadelphia: University of Pennsylvania Press.

Wilson, S.R., Stancavage, F.B., & Wise, L.L. (1981). *Synthesis of recent research on medical career decisions: A comparative study of two generations of physicians*. Palo Alto, CA: American Institutes for Research.

Steel, L.M., & Wise, L.L. (1977). *Designing a study of adult accomplishment and life quality*. Palo Alto, CA: American Institutes for Research.

Wise, L.L., McLaughlin, D.H., & Steel, L.J. (1977). *The Project TALENT data bank handbook*. Palo Alto, CA: American Institutes for Research.

Wise, L.L., McLaughlin, D.H., & Gilmartin, K.G. (1977). *The American citizen: Eleven years after high school, Volume II*. Palo Alto, CA: American Institutes for Research.

Gilmartin, K.G., McLaughlin, D.H., Wise, L.L., & Rossi, R.J. (1976). *Development of scientific careers: The high school years*. Palo Alto, CA: American Institutes for Research.

Rossi, R.J., Bartlett, W.B., Campbell, E.A., Wise, L.L., & McLaughlin, D.H. (1975). *Using the TALENT profiles in counseling: A supplement to the career data book*. Palo Alto, CA: American Institutes for Research.

Wilson, S.R., & Wise, L.L. (1975). *The American citizen: Eleven years after high school*. Palo Alto, CA: American Institutes for Research.

Presentations

Wise, L. L. (2015). *The Standards for Educational and Psychological Tests: Implications for Peer Review of State Assessments*. Workshop for state testing directors at the National Conference on Student Assessment, San Diego, CA.

Wise, L. L. (2015). *Psychometric Considerations for the Next Generation of Performance Assessment: What are the implications for state assessment programs?* Paper presented at the National Conference on Student Assessment, San Diego, CA.

Wise, L. L. (2015). *Educational Measurement: What lies ahead?* Presidential address to the annual meeting of the National Council on Measurement in Education. Chicago, IL.

Wise, L. L. (2015). *Psychometric Considerations for the Next Generation of Performance Assessment: Modeling, Dimensionality, and Weighting of Performance Task Scores*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Wise, L. L. (2015). *Test Design and Development Following the Standards for Educational and Psychological Testing*. Presentation to the annual meeting of the National Council on Measurement in Education. Chicago, IL.

Wise, L.L. (2013). *Different but Comparable: Good enough for government work?*. Paper presented at the annual meeting of the National Council on Measurement in Education. San Francisco, CA.

Wise, L.L. (2012). *Combining multiple indicators of achievement and growth*. Paper presented at the annual meeting of the National Council on Measurement in Education. Vancouver, BC.

Wise, L.L. (2012). *Prior linking efforts: The best laid plans* Paper presented at the annual meeting of the National Council on Measurement in Education. Vancouver, BC.

Wise, L.L. (2012). *Revising our Standards for Educational and Psychological Testing*. Paper presented at the Association of Test Publisher's Innovations in Test Design Conference. Palm Springs, CA.

Wise L. (April 2011). *Situating the Generalizability of Performance Assessments within a Validity Framework*. Paper presented at the 2011 Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Wise, L. (April 2011). *Aggregating Results from Through-Course Assessments*. Paper presented at the 2011 Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Wise, L. (April 2011). *Aggregating Results from Through-Course Assessments*. Paper presented at the National Conference on Student Assessment, Orlando, FL.

Wise, L. (April 2011). *The Evolving U.S. Educational System: How Can I-O Psychology Contribute?* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Chicago, IL.

Wise, L. (June 2011). Update on Revision of the *Standards for Educational and Psychological Testing*. Paper presented at the National Conference on Student Assessment, Orlando, FL.

Wise, L. (August 2011). Update on Revision of the *Standards for Educational and Psychological Testing*. Paper presented at the annual convention of the American Psychological Association, Washington, DC.

Wise, L.L. (2009) *Revising our Test Standards: Issues with Increased Use of Tests for Accountability*. Presentation at the 2009 annual meeting of the American Psychological Association, Toronto, Canada.

Wise, L.L. (2009, June) *Revising our Test Standards*. Presentation to the CCSSO National Conference on Student Assessment, Los Angeles, CA.

- Wise, L.L. (2009, April) *Revising our Test Standards: Issues for Work Place Testing*. Presentation at the 2009 annual meeting of the American Educational Research Association, San Diego, CA.
- Wise, L.L. (2008). *Validating Indicators of College Preparedness: Ready or Not?* Presentation to the 2008 Reidy Interactive Lecture Series. Portsmouth, NH
- Wise, L.L. (September 2008). *State Assessments Today: What State are We In?* Presentation to the Conference on Educational Testing in America: State Assessment Achievement Gaps, Federal Policy and Innovations. Washington, DC.
- Wise, L.L. (June 2008). *Strengthening K-12 Accountability Systems: What could be better than peer review?*. Presentation to the CCSSO National Conference on Student Assessment, Orlando, FL.
- Wise, L.L., & Plake, B. (June 2008). *Procedures for Revising the Test Standards*. Presentation to the CCSSO National Conference on Student Assessment, Orlando, FL.
- Wise, L.L., & Rui, N. (2008, March). *Computing and Communicating Test Accuracy for High-Stakes Decisions*. Paper presentation at the 2008 annual meeting of the National Council on Measurement in Education, New York City, NY.
- Wise, L.L. (2007). *Vertical Alignment*. Paper presented at the CCSSO Large Scale Assessment Conference, Nashville, TN.
- Shen, X. (Stanford University; former HumRRO intern), Wise, L.L., and Becker, D.E. (2006, April). Analysis of School Effects. In L. Roberts (Chair), *Evaluating the Impact of a High-School Graduation Test*. Symposium conducted at the annual conference of the American Educational Research Association, San Francisco.
- Wang, X. , Wise, L.L., Becker, D.E., and Taylor, L.R. (2006, April). Comparison of CAHSEE Content Standards to Graduation and High-School Accountability Tests Used in Other States. In L. Roberts (Chair), *Evaluating the Impact of a High-School Graduation Test*. Symposium conducted at the annual conference of the American Educational Research Association, San Francisco.
- Wise, L. (Discussant) (2006, May). *Expanding our influence: How I –O psychologists can improve education*. Practice Forum conducted at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.
- Wise, L.L. and Becker, D.E. (2006, April). Analysis of Results from Administrations of the California High School Exit Exam (CAHSEE). In L. Roberts (Chair), *Evaluating the Impact of a High-School Graduation Test*. Symposium conducted at the annual conference of the American Educational Research Association, San Francisco.
- Wise, L. W. (2004). *Independent evaluation of the California High School Exit Exam (CAHSEE)*. Presentation at the American Educational Research Association Annual Meeting, San Diego, CA.
- Wise, L. W. (2004). *Improving scientific research in education: Recent activities of the National Research Council*. Presidentially Invited Symposium conducted at the American Educational Research Association Annual Meeting, San Diego, CA.

Wise, Laress L. (2004). Vertically articulated content standards. Presentation for the Reidy Interactive Lecture Series. Nashua, NH.

Wise, Laress L. (2004). Meeting Alignment challenges: Analyzing vertical alignment. Paper presented at the CCSSO Large Scale Assessment Conference. Boston, MA.

Wise, Laress L. (2004). Independent Evaluation of the CAHSEE: Update on Evaluation Findings and Recommendations. Presentation to the California State Board of Education, Sacramento, CA.

Wise, Laress L. (2004). Debra P. v. Turlington and CAHSEE: Those who do not study history are doomed to repeat it: Presidential Invited Session. Annual Meeting of the American Educational Research Association. San Diego, CA.

Wise, L. L., Floden, R.E., Dickersin, K. & Schneider, B.L. (2004). Improving Scientific Research in Education: Recent Activities of the National Research Council. Presidential Invited Session. Annual Meeting of the American Educational Research Association. San Diego, CA.

Wise, L.L. (2001). *Building a selection test battery: Validity, fairness, and other tradeoffs*. Presentation to the Personnel Testing Council/Metropolitan Washington. Washington, DC: National Research Council.

Wise, L.L. (2001). Validity, Fairness, and other Tradeoffs. Presentation to the Personnel Testing Council/Metropolitan Washington. Washington, DC: National Research Council.

Wise, L.L. (1999). *How far should NAEP go in serving an interpretive function?* Presentation to the Forum on NAEP Design: 2000-2010. Washington, DC: National Research Council.

Wise, L.L. & Hauser, R.M. (1999). Evaluation of the Voluntary National Tests. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Waugh, G.W., Wise, L.L., Quartetti, D.A., & Ramos, R.A. (1999). Validation of the air traffic controller predictor tests. In R.A. Ramos, (Chair), *Air traffic selection and training project*. Symposium conducted at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Inc., Atlanta, GA.

Wise, L.L., Quartetti, D.A. (HumRRO), Kieckhafer, W.F. (RGI), & Houston, J.S. (PDRI). (1999). Development of air traffic controller predictor battery. In R.A. Ramos (Chair), *Air traffic selection and training project*. Symposium conducted at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Inc., Atlanta, GA.

Wise, L.L. (1999, April). How far should NAEP go in serving an interpretive function? Presentation to the Forum on NAEP Design: 2000-2010. Washington, DC: National Research Council.

Wise, L.W. (1999). Test-taking motivation as persistence: Its effect on item and test performance. Paper presented at the Festshift for William Meredith, Berkeley, CA.

Wise, L.L., Quartetti, D.A., Kieckhafer, W.F., & Houston, J.S. (1999). Development of air traffic controller predictor battery. In R.A. Ramos (Chair), *Air traffic selection and training project*.

Symposium conducted at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Inc., Atlanta, GA.

Wise, L.L. (1998, April). Relationship of Kentucky High School Assessment Scores to Predictors. In R.G. Hoffman (Chair), Symposium conducted at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Wise, L.L. (1998, April). Discussant in Statistical support for CAT implementation. Paper Session at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Wise, L.L. (1997). Career directions beyond the dissertation: R&D "think tanks." Presenter at invited symposium, Career directions in measurement: Beyond the dissertation, sponsored by NCME Graduate Student Issues Committee and co-sponsored by AERA Division D, at the 1997 National Council on Measurement in Education Annual Meeting, Chicago, IL.

Wise, L.L. (1996, April). A persistence model of motivation and test performance. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Curran L.T., & Wise, L.L. (1994). Enlistment processing changes. Paper presented at the annual meeting of the International Military Testing Association, The Netherlands.

Welsh, J.R., & Wise, L.L. (1994). The stability of the Mantel-Haenszel odds ratio. Paper presented at the annual meeting of the International Military Testing Association, The Netherlands.

Wise, L.L., & Welsh, J.R. (November 1993). Order adjustment and cross-correlation in ASVAB test form development. Paper presented at the annual meeting of the International Military Testing Association, Williamsburg, VA.

Wise, L.L., & Wall, J.E. (1993, November). Plan for evaluating the ASVAB career exploration program. Paper presented at the annual meeting of the International Military Testing Association, Williamsburg, VA.

Wise, L.L., & Curran, L.T. (1993, August). Introducing the new ASVAB: Recommendations and decisions for change. Paper presented at the annual meeting of the American Psychological Association, Toronto, Canada.

Wise, L.L. (1993, April). Scoring rubrics for performance tests: Lessons learned from job performance assessment in the military. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA

Wise, L.L. (1993, April). Test form accuracy. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Wise, L.L. (1992, April). Lessons learned from military performance assessment. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Wise, L.L. (1991, October). Overview of the ASVAB revision process. Paper presented at the annual meeting of the Military Testing Association, San Antonio, TX.

Wise, L.L., & McDaniel, M.A. (1991). Cognitive factors in the Armed Services Vocational Aptitude Battery and the General Aptitude Test Battery. Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA. Wise, L.L., Chia, W.J., & Park, R.K. (1989). Effects of item position on IRT parameter estimates and item statistics. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Arabian, J.M., Wise, L.L., & Szenas, P.L. (1988). Setting performance standards for Army enlisted jobs. Paper presented at the annual meeting of the American Psychological Association, New Orleans, LA.

Szenas, P.L., Wise, L.L., & Arabian, J.M. (1988). Combining individual standards into an overall standard: Modeling the judgement process and investigating differences among judges. Paper presented at the annual meeting of the American Psychological Association, New Orleans, LA.

Wise, L.L., McHenry, J.J., & Campbell, J.P. (1987). Matching skills and traits to job requirements: Results from Project A. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Wise, L.L. (1987). Differential item difficulty indicators in small samples. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Wise, L.L., McHenry, J.J., Rossmeissl, P.G., & Oppler, S.H. (1986). ASVAB validities using improved job performance measures. Paper presented at the annual meeting of the Military Testing Association, Mystic, CT.

Wise, L.L., Campbell, J.P., McHenry, J.J., & Hanser, L.M. (1986). A latent structure model of job performance factors. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Wise, L.L. (1986). Latent trait models for partially speeded tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wise, L.L., & Mitchell, K.J. (1985). Development of an index of maximum validity increment for new predictor measures. Paper presented at the annual meeting of the American Psychological Association, Los Angeles, CA.

Wise, L.L., & Wilson, S.R. (1982). Test item calibration. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Wise, L.L., & McLaughlin, D.H. (1981). Survey data enhancement. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Wise, L.L., Wilson, S.R., & Stancavage, F.B. (1980). The development of medical practice and residence value scales that distinguish physicians in different specialties and practice locations. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Wise, L.L., & Steel, L.M. (1979). The effects of school quality on student's knowledge and skills. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wise, L.L. (1979). Long-term consequences of sex differences in high school mathematics education. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Steel, L.M., & Wise, L.L. (1979). Origins of sex differences in high school mathematics achievement and participation. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wise, L.L. (1978). The role of mathematics in women's career development. Paper presented at the annual meeting of the American Psychological Association, Toronto, Canada.

**UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA**

AMERICAN EDUCATIONAL RESEARCH)	
ASSOCIATION, INC., AMERICAN)	
PSYCHOLOGICAL ASSOCIATION, INC.,)	
and NATIONAL COUNCIL ON)	
MEASUREMENT IN EDUCATION, INC.,)	Civil Action No. 1:14-cv-00857-TSC-DAR
)	
Plaintiffs,)	DECLARATION OF WAYNE
)	CAMARA IN SUPPORT OF
v.)	PLAINTIFFS' MOTION FOR
)	SUMMARY JUDGMENT AND ENTRY
PUBLIC.RESOURCE.ORG, INC.,)	OF A PERMANENT INJUNCTION
)	
Defendant.)	
)	

I, WAYNE J. CAMARA, declare:

1. I am the Senior Vice President, Research at ACT. My company produces and publishes the ACT[®] college readiness assessment — a college admissions and placement test taken millions of high school graduates every year. ACT also offers comprehensive assessment, research, information, and program management services to support education and workforce development. As the Senior Vice President of Research, I am responsible for all research and evidence related to the design, development, use, and validation of our assessments and programs. In my position, I serve on the Senior Leadership Team and manage over 110 researchers.

2. I submit this Declaration in support of the motion of the American Educational Research Association, Inc. (“AERA”), the American Psychological Association, Inc. (“APA”), and the National Council on Measurement in Education, Inc. (“NCME”) (collectively, “Plaintiffs” or “Sponsoring Organizations”) for summary judgment and the entry of a permanent injunction.

3. Prior to working at ACT, I worked at The College Board, where I held the

positions of Vice President, Research and Development (July, 2000 – September, 2013), Executive Director, Office of Research and Development (March, 1997 – June, 2000), and Research Scientist (September, 1994 – February, 1997).

4. Before working at The College Board, I worked for APA in the positions of Assistant Executive Director for Scientific Affairs and Executive Director of Science (1992-1994), Director, Scientific Affairs (February, 1989 – August, 1992), and Testing and Assessment Officer (November, 1987 – January, 1989.) During my time at APA, I also served as the Project Director for the revision of the 1985 edition of the *Standards for Educational and Psychological Testing* published in 1999 (the “1999 Standards”). In 1997, I was elected to APA’s Council of Representatives, and I served on the Council from 1997-2003. In April, 2012, I was elected to the AERA Council, serving from April, 2012 to April, 2015 as Vice President for Division D. I was also elected to NCME’s Board of Directors, serving on the Board from 2002-2005 and 2009-2012, and served as NCME’s President from 2010-2011. Additionally, I have served on the Management Committee for the Standards from 2005-2015.

5. My curriculum vitae is attached to this Declaration as Exhibit 1.

6. I have written extensively on the Standards, as well as other professional and technical guidelines which relate to educational and industrial testing and assessment, including journal articles, book chapters, and paper presentations at national conferences.

7. In 1954, APA prepared and published the “Technical Recommendations for Psychological Tests and Diagnostic Techniques.” In 1955, AERA and NCME prepared and published a companion document entitled, “Technical Recommendations for Achievement Tests.” Subsequently, a joint committee of the three organizations modified, revised, and consolidated the two documents into the first Joint Standards. Beginning with the 1966 revision,

the Sponsoring Organizations collaborated in developing the “Joint Standards” (or simply, the “Standards”). Each subsequent revision of the Standards has been careful to note that it is a revision and update of the prior version.

8. Beginning in the mid-1950s, the Sponsoring Organizations formed and periodically reconstituted a committee of highly trained and experienced experts in psychological and educational assessment, charged with the initial development of the Technical Recommendations and then each subsequent revision of the (renamed) Standards. These committees were formed by the Sponsoring Organizations’ Presidents (or their designees), who would meet and jointly agree on the membership. Often a chair or co-chairs of these committees were selected by joint agreement. Beginning with the 1966 version of the Standards, this committee became referred to as the “Joint Committee.”

9. Financial and operational oversight for the Standards’ revisions, promotion, distribution, and for the sale of the 1999 and 2014 Standards has been undertaken by a periodically reconstituted Management Committee, comprised of designees of the three Sponsoring Organizations. As noted above, I served on this Management Committee from 2005-2015.

10. All members of the Joint Committee(s) and the Management Committee(s) are *unpaid* volunteers. The expenses associated with the ongoing development and publication of the Standards include travel and lodging expenses (for the Joint Committee and Management Committee members), support staff time, printing and shipment of bound volumes, and advertising costs.

11. From the time of their initial creation to the present, the preparation of and periodic revisions to the Standards entail intensive labor and considerable cross-disciplinary

expertise. Each time the Standards are revised, the Sponsoring Organizations select and arrange for meetings of the leading authorities in psychological and educational assessments (known as the Joint Committee). During these meetings, certain Standards are combined, pared down, and/or augmented, others are deleted altogether, and some are created as whole new individual Standards. The 1999 version of the Standards is nearly 200 pages, took more than five years to complete, and is the result of work put in by the Joint Committee to generate a set of best practices on educational and psychological testing that are respected and relied upon by leaders in their fields.

12. Draft revisions of the 1985 Standards, for what became the 1999 Standards, were widely distributed for public review and comment during the revision process. The Joint Committee received thousands of pages of comments and proposed text revisions from: the membership of the Sponsoring Organizations, scientific, professional, trade and advocacy groups, credentialing boards, state and federal government agencies, test publishers and developers, and academic institutions. While the Joint Committee reviewed and took under advisement these helpful comments, the final language of the 1999 Standards was a product of the Joint Committee members. When the 1985 Standards were revised, more than half the content of the 1999 Standards resulted from newly written prose of the Joint Committee.

13. The Standards originally were created as principles and guidelines – a set of best practices to improve professional practice in testing and assessment across multiple settings, including education and various areas of psychology. The Standards can and should be used as a recommended course of action in the sound and ethical development and use of tests, and also to evaluate the quality of tests and testing practices. Additionally, an essential component of responsible professional practice is maintaining technical competence. Many professional

associations also have developed standards and principles of technical practice in assessment. The Sponsoring Organizations' Standards have been and still are used for this purpose.

14. The Standards, however, are not simply intended for members of the Sponsoring Organizations, AERA, APA, and NCME. The intended audience of the Standards is broad and cuts across audiences with varying backgrounds and different training. For example, the Standards also are intended to guide test developers, sponsors, publishers, and users by providing criteria for the evaluation of tests, testing practices, and the effects of test use. Test user standards refer to those standards that help test users decide how to choose certain tests, interpret scores, or make decisions based on tests results. Test users include clinical or industrial psychologists, research directors, school psychologists, counselors, employment supervisors, teachers, and various administrators who select or interpret tests for their organizations. There is no mechanism, however, to enforce compliance with the Standards on the part of the test developer or test user. The Standards, moreover, do not attempt to provide psychometric answers to policy or legal questions.

15. The Standards promote the development of high quality tests and the sound use of results from such tests. Without such high quality standards, tests might produce scores that are not defensible or accurate, not an adequate reflection of the characteristic they were intended to measure, and not fair to the person tested. Consequently, decisions about individuals made with such test scores would be no better, or even worse, than those made with no test score information at all. Thus, the Standards help to ensure that measures of student achievement are relevant, that admissions decisions are fair, that employment hiring and professional credentialing result in qualified individuals being selected, and patients with psychological needs are diagnosed properly and treated accordingly. Quality tests protect the public from harmful

decision making and provide opportunities for education and employment that are fair to all who seek them.

16. The Standards apply broadly to a wide range of standardized instruments and procedures that sample an individual's behavior, including tests, assessments, inventories, scales, and other testing vehicles. The Standards apply equally to standardized multiple-choice tests, performance assessments (including tests comprised of only open-ended essays), and hands-on assessments or simulations. The main exceptions are that the Standards do not apply to unstandardized questionnaires (*e.g.*, unstructured behavioral checklists or observational forms), teacher-made tests, and subjective decision processes (*e.g.*, a teacher's evaluation of students' classroom participation over the course of a semester).

17. The Standards have been used as a source in developing testing guidelines for such activities as college admissions, personnel selection, test translations, test user qualifications, and computer-based testing. The Standards also have been widely cited to address technical, professional, and operational norms for all forms of assessments that are professionally developed and used in a variety of settings. The Standards additionally provide a valuable public service to state and federal governments as they voluntarily choose to use them. For instance, each testing company, when submitting proposals for testing administration, instead of relying on a patchwork of local, or even individual and proprietary, testing design and implementation criteria, may rely instead on the Sponsoring Organizations' Standards to afford the best guidance for testing and assessment practices.

18. The Standards were not created or updated to serve as a legally binding document, in response to an expressed governmental or regulatory need, nor in response to any legislative action or judicial decision. However, the Standards have been cited in judicial decisions related

to the proper use and evidence for assessment, as well as by state and federal legislators. These citations in judicial decisions and during legislative deliberations occurred without any lobbying by the Plaintiffs.

19. The Sponsoring Organizations do not keep any of the revenues generated from the sales of the Standards. Rather, the income from these sales is used by the Sponsoring Organizations to offset their development and production costs and to generate funds for subsequent revisions. This allows the Sponsoring Organizations to develop up-to-date, high quality Standards that otherwise would not be developed due to the time and effort that goes into producing them.

20. At one time, funding for the Standards revision process from third party sources (*e.g.*, governmental agencies, foundations, other associations interested in testing and assessment issues, etc.) was raised as a consideration. However, this option was not seriously explored as the potential conflicts of interest in doing so left the Sponsoring Organizations to conclude that the Standards revisions should be self-funding – that is, from the sale of prior editions of the Standards.

21. In late 2013 and early 2014, the Sponsoring Organizations became aware that the 1999 Standards had been posted on the Internet without their authorization, and that psychology students were obtaining free copies from the posting source. Upon further investigation, the Sponsoring Organizations discovered that Public.Resource.Org, Inc. (“Public Resource”) was the source of the online posting. Accompanying this Declaration as Exhibit MMM is a true copy of a thread of emails exchanged among Laurie Wise, Suzanne Lane, David Frisbie, Jerry Sroufe, Marianne Ernesto, Barbara Plake, and myself¹ sent between December 16, 2013 and February 4,

¹ Laurie Wise is the Immediate Past President of NCME and was serving as President of NCME at the time of the email, Suzanne Lane is a member of the Standards Management Committee; David Frisbie also is a member of the

2014, discussing Public Resource's posting of the 1999 Standards on the Internet, and marked as Exhibit 1185 during my deposition.

22. Past harm to the Sponsoring Organizations from Public Resource's activities includes a lack of greater funding that otherwise would have been available for the update of the Sponsoring Organizations' Standards from the 1999 to the 2014 versions, due to the reduced volume of sales of the 1999 Standards.

23. Should Public Resource's infringement be allowed to continue, the harm to the Sponsoring Organizations, and public at large who rely on the preparation and administration of valid, fair and reliable tests, includes: (i) uncontrolled publication of the 1999 Standards without any notice that those guidelines have been replaced by the 2014 Standards; (ii) future unquantifiable loss of revenue from sales of authorized copies of the 1999 Standards (with proper notice that they are no longer the current version) and the 2014 Standards; and (iii) lack of funding for future revisions of the 2014 Standards and beyond.

24. Due to the small membership size of NCME, and the relative minor portion of the membership of AERA and APA who devote their careers to testing and assessment, it is highly unlikely that the members of the Sponsoring Organizations will vote for a dues increase to fund future Standards revision efforts if Public Resource successfully defends this case and is allowed to post the Standards online for the public to download or print for free. As a result, the Sponsoring Organizations would likely abandon their practice of periodically updating the Standards and there would be an absence of any authoritative and independent source of sound guidance relating to the development, use, and evaluation of psychological and educational tests.

Standards Management Committee; Jerry Sroufe is the Director of Government Relations at AERA, Marianne Ernesto is the Director, Testing and Assessment, at APA, and Barbara Plake was Laurie Wise's co-chair of the Joint Committee for the revision of the 1999 Standards, which ultimately were published in 2014.

Dated: December 8, 2015



Wayne J. Camara

EXHIBIT 1

WAYNE J. CAMARA

OFFICE:

ACT
500 ACT Drive
Iowa City, IA 52243-0168
Tel (319) 337-1869
wayne.camara@act.org

HOME:

81 Lewis St.,
Marion, MA 02738

EDUCATION:

Ph.D. 1986 University of Illinois at Urbana, Psychology.
Educational Measurement
Cognate: Industrial and Organizational Psychology

C.A.G.S. 1982 Rhode Island College (School Psychology), Providence, R.I.

M.A. 1980 Rhode Island College (Educational Measurement),
Providence, R.I.

B.A. 1978 University of Massachusetts (Psychology/Education), N. Dartmouth, MA

PROFESSIONAL EXPERIENCE:

ACT, Iowa City, IA

Senior Vice President, Research (September 2013 -)

Oversees research departments across education and workforce assessments and services related to research, psychometrics, data reporting, statistical analysis, policy research, survey development, and industrial psychology services (e.g., job profiling). Manage a staff of over 125 professional staff. Serves on ACT's strategic leadership team and is responsible for shaping and guiding organizational direction and planning, as well as representing the organization with external audiences and stakeholders in areas including accountability, research, admissions testing, etc. Member of the Executive Leadership Team and business sponsor on a range of technology and development projects.

The College Board, New York, NY

Vice President, Research and Development (July, 2000 – September 2013)

Executive Director, Office of Research and Development (March, 1997 - Present)

Research Scientist (Sept. 1994 - Feb., 1997)

Was responsible for all research, standards and alignment services, psychometric and assessment development activities at the College Board, including design and implementation of R&D activities that support College Board assessment programs (SAT, PSAT/NMSAT, AP, CLEP, Subject Tests, Accuplacer, SpringBoard, etc.). Managed a staff of approximately 75 professionals across several units and locations: Research, Statistics and Psychometrics, Test Development, Analysis and Reporting, and Standards Alignment. Responsible for policy research, outreach with state assessment directors, higher educational institutions, state Boards of Education and other policy and governance bodies. Coordinate product planning and business planning for new assessments and enhancements to current assessments. Responsible for several external advisory panels and test development committees. Responsible for reporting SAT aggregate results to institutions, reviewing all items and final forms of the SAT, and other

operational work related to assessment development and delivery. Serves as a spokesperson for the College Board on technical and assessment policy discussions with the media, policymakers (e.g., testimony), institutions and other key stakeholders. Works with states, districts, policy makers and higher educational systems, to provide data, analyses and information concerning student achievement and college readiness. Directed data release process, guidelines and approvals.

Project manager for development of the New SAT and represents the College Board on issues of test development and research with universities, higher educational associations, states and districts, academic associations and other groups. Responsible for hiring / management of vendors and academicians to implement research, review test forms and items, and prototype development.

Specific areas of research include validity of admissions measures, evaluation of educational programs, include effects of accommodations and extended time for examines with disabilities, meta-analysis of SAT validity, grade inflation trends, and development of new constructs and measures relevant to an expanded predictor - criterion space.

Selected development efforts: (1) Conceived, developed and conducted research resulting in AP Potential which is increased access to AP courses by identifying students with potential for success; (2) 2005 SAT redesign with writing; (3) Implementation of ECD and AP Redesign work in selected courses; (4) Research and transition plan to move AP, SAT and PSAT from formula scoring to rights only scoring; (5) Design on AP Portfolio and through course pilot; (6) Accuplacer diagnostic tests and replatforming; (7) Plan to migrate most research and selected psychometric operational work to the CB from vendors; and (8) Design of CLEP-testlet assessment.

AMERICAN PSYCHOLOGICAL ASSOCIATION, Washington, D.C.

*Assistant Executive Director for Scientific Affairs and Executive Director of Science (1992 -1994.) Director, Scientific Affairs (February 1989 - Aug., 1992)
Testing and Assessment Officer (Dec. 1987 - Jan. 1989)*

Project Director for the Revision of the Standards for Educational and Psychological Testing and Assessment. Managed the technical committee, various technical panels, a financial management committee and an executive committee comprised of the Presidents of APA, AERA, and NCME.

Coordinated and developed all association policies and guidelines in areas of scientific affairs, scientific misconduct, research funding, and testing and assessment. Major area of responsibilities in measurement and assessment included: (a) monitoring scientific and technical advances; (b) educating policy makers, the public, the media, and other professionals (e.g., employers, educators) of the relevance and appropriate applications of assessment; (c) developing technical guidance and policy statements that address new and emerging areas, reflecting both the scientific and professional consensus in assessment; (d) working collaboratively with other professional associations, advocacy groups, and governmental agencies; and (e) testimony and advocacy on the efficacy of behavioral science.

Directed APA involvement in numerous assessment issues at the national level: SCANS, Americans with Disabilities Act, national education standards, industry-based skills standards, Civil Rights Act of 1991, efficacy of clinical assessment, integrity testing, and test-based accountability initiatives. Assisted in developing amicus briefs for Supreme Court, informing policymakers, media, and the public of technical advances in assessment (e.g., validation strategies, computer-based and interactive assessments, implications of fairness and utility analyses, etc.) and behavioral science research more broadly.

GEORGE WASHINGTON UNIVERSITY, Washington DC

Adjunct Professor of Administrative Sciences and College of Business (1988 - 1994)

Taught graduate seminars in training, performance evaluation, personnel selection, and organizational behavior. Served on several doctoral dissertations in I-O psychology.

HUMAN RESOURCES RESEARCH ORGANIZATION, Alexandria, VA,

Senior Scientist (Feb. 1987 - November 1987), Research Scientist (August 1985 - Feb. 1987)

Conducted research and managed grants and proposal development in areas of job analysis, competency modeling, military testing, training, and personnel selection. Projects including:

Investigated the utility of algorithms used in computer-based job classification systems employed by each branch of the military service. Developed a crosswalk between military occupations in each service branch and civilian occupations.

Project Director and Principal Investigator for contracts funded by the Assistant Secretary of Defense and the Navy Personnel Research and Development Center to conduct a longitudinal evaluation of the impact of military training and service on subsequent employment/social success of low aptitude youth enlisted in the military.

Developed training and career development system for first-line civilian supervisors in the U.S. Army. Provided recommendations for the career development and training of future and incumbent Army civilian first-line supervisors.

Developed training evaluation instruments and conducted evaluation of counselor training in the use and interpretation of the ASVAB.

Managed and conducted several job analysis projects for military and civilian occupations with the Department of Defense.

PERSONNEL SERVICES OFFICE, UNIVERSITY OF ILLINOIS, Champaign, IL

Human Resources Consultant (1983-85), Illinois State Civil Service

Designed and managed R&D projects including the development of a computerized adaptive screening measure to optimize the matching of jobs and applicants of Civil Service positions. Conducted a large-scale job analysis of 70 professional and technical job classifications. Used multiple-rater, multiple-method job analyses and applied generalizability theory to interpret findings. Performed validation studies of existing civil service exams.

DEPARTMENT OF PSYCHOLOGY, UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Teaching Assistant and Academic Advisor (1983-85). Research Assistant (1984-85).

COLLEGE OF EDUCATION, UNIVERSITY OF ILLINOIS

Psychological Evaluator (1983-84). Administered, scored, and interpreted a variety of psychological and cognitive measures.

BRISTOL COMMUNITY COLLEGE, Fall River, MA

Lecturer (Spring 1980 -1982), Psychology and Education

WEST BRIDGEWATER PUBLIC SCHOOLS, West Bridgewater, MA
School Psychologist (1979-82)

Chairperson on team evaluations and reviews. Representative for the school district in out-of-district placements, conferences, and regional and state planning meetings.

Psychological testing – psychodiagnostic and learning assessment - individual IQ tests, projective testing, special abilities testing, including two years of clinical supervision. Developed detailed assessment and remediation plans for over 150 students.

PROFESSIONAL AFFILIATIONS:

American Psychological Association - Elected Fellow, 1994
Division of Educational Psychology
Division of Evaluation, Measurement, and Statistics – Elected Fellow, 2002
Division of General Psychology – Elected Fellow, 1994
Division of Military Psychology
Division of International Psychology – Elected Fellow, 2009
American Psychological Society, Elected Fellow 2007
American Educational Research Association – Elected Fellow 2008
International Association of Applied Psychology
International Test Commission 1997
National Association of Collegiate Admissions Counselors
National Council for Measurement in Education
New York Academy of Sciences – Elected, 2002
Personnel Testing Council of Metropolitan Washington
Society for Industrial/Organizational Psychology – Elected Fellow, 1999

RELATED PROFESSIONAL ACTIVITIES AND AWARDS:

American Educational Research Association, Division D (Measurement and Research Methodology);
Co-Chair, Annual Convention Program, 1998.

American Psychological Association: Appointed APA Member of the Joint Committee on Testing Practices, 1997-2000; Elected to Council of Representatives, 1997-2000; Member of Joint Science and Practice Integration Task Force, 1998; Member CODAPAR, 2004-2007.

Division of Evaluation, Measurement and Statistics, Member-at-Large, 1997-2000; Chair;
Professional Affairs, 1995-96, Member, Program Committee, 1992, 1994.

Division of Military Psychology: Chair, Program Committee, 1988.

Division of General Psychology: Chair, Membership Committee, 1996; Member-at-Large, 2002-2005.

Associate Editor, Journal of Occupational Health Psychology, 1996 – 1999; Journal of Experimental Psychology – Applied, 2001 – 2007; Advisory Editor, Journal of Educational Measurement 2008 – , Educational Measurement: Issues and Practice, 2012 - .

Audit team, Psychometric and measurement graduate program, University of North Carolina at Greensboro, spring, 2006.

Author of numerous technical and policy statements approved by the American Psychological Association (e.g., Statement on the disclosure of test data, Statement on the Golden Rule, Resolution on a separate directorate for behavioral sciences at NSF).

Author, test reviews on numerous employment, organizational and career tests for *Buros Mental Measurements Yearbook* 1994, 1996, 1997, 1998, 2000, 2005, 2006, 2008.

Award for Distinguished Service Contributions, the Society of Industrial and Organizational Psychology, 2004.

Award for Professional Contributions and Service to Testing, Association of Test Publishers, 2014.

Award (to staff unit) for Dissemination of Educational Measurement Concepts to the Public, Code of Fair Testing Practices in Education National Council on Measurement in Education, 1989

Award for New Product Development, Testlet Design for CLEP, Educational Testing Service, 1998.

Board of Advisors, Center for Enrollment Research Policy and Practice, University of Southern California, 2008 –

Council of Chief State School Officers, Technical Issues in Large Scale Assessment (TILSA), 2014 -

Common Core State Standards – Assisted in development and policy oversight in joint effort led by CCSSO and National Governors Association (2009-10).

Editorial Board, *International Journal of Selection and Assessment* (2001 - 06), *Military Psychologist* (2002 - 07), *Journal of Experimental Psychology: Applied* (2001 - 07), *Educational Measurement: Issues and Practice* (2010 – current), NCME Edited Book Series (2014 – current).

Expert in judicial and regulatory proceedings involving cognitive ability testing, accommodations and score comparability in admissions testing, personality testing and disparate impact, job analysis and recruitment practices, affirmative action (*Gratz v. Bollinger*), and copyright infringement on the *Standards for Educational and Psychological Testing*.

Independent Consultant (selected list), American Council on Education, Goodyear Corporation, American Waterways, Federal Reserve Bank of New York, City University of New York, Maryland State Departments of Education, Army Research Institute, American Institute for Research, US DOE, Tennessee Department of Education, PSI, Wonderlic Inc., employment and labor attorneys and several other organizations in areas of employment testing, educational evaluation, college readiness and standard setting, performance appraisal systems, and survey research.

Journal Reviewer: *American Psychologist*; *Educational Measurement: Issues and Practice*; *Educational Researcher*; *Personnel Psychology*; *Psychology, Public Policy and the Law*; *Journal of Occupational Health Psychology*; *Journal of Educational Measurement*, *Applied Educational Measurement*, *Journal of Educational Measurement*, *Educational Measurement: Issues and Practice*, *Journal of Applied Psychology*, *Human Factors*, *Military Psychologist*, etc.

Media experience: Appeared on national and local television and radio (CNN, Good Morning America, BBC, PBS) to discuss the use Civil Rights Act, ADA, personality testing and admissions testing; Frequently quoted in major newspaper stories involving testing, 1992- Present.

Member of International Standards Organization (ISO) Working Group on International Standards on Psychological testing (ANSI), 2007 – 2010.

National Academy of Sciences, Panelist and participant in workshops sponsored by the Board of Testing and Assessment on School-to-Work (1997-98), Collegiate Admissions testing, and Accommodations and flagging test scores for disabled test takers (1997, 2003).

National Council on Education in Measurement,
Chair, Professional Responsibilities Committee, 1996 - 2000.
Chair, Career Award Committees, 2015-2016.
Fund Development Committee, 2013-2016

Office of Educational Research and Improvement, Technical Review Committee for grants associated with the National Assessment of Educational Progress, 1992-1997.

Society for Industrial and Organizational Psychology: Member of Executive Committee, 1988- 2003; Chair, External Affairs Committee, 1993-95; Chair, Awards Committee, 1991-93; Chair, Membership Committee, 1988-91; Membership Committee, 1987-88; Program Committee 1986-87; 1998-99; Fellowship Committee 2007-10; and Distinguished Service Award 2011-13. Designed membership survey and developed first SIOP membership directory.

Standards for Educational and Psychological Testing, AERA, APA and NCME. Staff Director (1992-94); Chair, Management Committee (2005-2015).

Standard Setting Approaches and Policy Capturing for College and Career Readiness (Consultation to several states) (2010-current).

- NAEP linkages and alignment studies with SAT and Accuplacer (2011-12) and ACT, Explore, Compass (2014-15).
- STARR end-of-course examinations, Texas Educational Agency (2012)
- End-of-course tests, Tennessee Department of Education (2011)
- Achieve Inc. Algebra II examination (2008-10).
- New York State (2012-13, through College Board contract).
- Wyoming, Department of Education (2014, ACT contract).
- South Carolina, Department of Education (2015, ACT contract).

Technical or Scientific Advisory Committee Member:

- Advisory Panel and Steering Panel, Department of Labor-OERI effort to develop assessments to measure competencies from the Secretary's Commission on Achieving Necessary Skills (SCANS), ACT, Iowa City, IA, 1992-94.
- American Association of Medical Colleges, Blue Ribbon Technical Panel on additional measures for admission to medical colleges, 2012 - 2015.
- American Diploma Project, Multi-state Algebra assessments, Research Alliance supported by Achieve for 16 states, 2007 – 2010.
- American Institute of Certified Public Accountants, Psychometric Oversight Committee, 2011 – 2014.
- Army Research Institute for the Behavioral and Social Sciences, Chair Scientific Review Panel on Selection and Classification Program, 2003; Panel member of the technical advisory panel on ABLE, 2001-02.
- Congressional Office of Technology Assessment. Technical assistance for study on integrity testing in employment settings, 1989-90, and study on performance assessments in school testing, 1991.
- Delaware State Education Department, Chair TAC on Race to the Top 2011-2013.

- Department of Defense, ASVAB testing program, 2000-2008, (chair 2002-2008).
- International Standards Organization, ISO Standard 10667 (organizational assessment), U.S. team on international development committee 2009-2012.
- Law School Admissions Council (chair), Technical Audit Team 2009.
- Metametrics, Technical Advisory Committee, 2013-2016.
- National Assessment of Educational Progress (NAEP), College freshmen technical panel, 2009 – 2010; Technical Advisory Committee on Standard Setting (Writing), 2010-2014; Advisory panel on survey of higher educational institutions on use of assessments for College Readiness and Placement, 2011-12.
- NCAA Data and Analysis Research Group, 2005-2008.
- Nebraska State Department of Education (TAC) 2008 – 2013.
- PARCC – Technical Advisory Committee, 2010 – 2013.
- Pearson Test of English, Scientific Advisory Committee, 2009 – 2013.
- Pennsylvania State Department of Education (TAC) 2003 - current
- Psychological Services Inc., employment-certification testing, Scientific Advisory Board, 2011 - current
- Texas State Department of Education (TAC) 2011 - current
- Technical Advisor Reporting Jointly to Texas Educational Authority and Texas Higher Education Coordinating Board 2008.
- U.S. Department of Education, National Advisory Technical Panel on NCLB Reporting, 2008-11.
- U.S. Department of Labor, Technical advisor, National Job Analysis and Skills Assessment, 1993-96.

U.S. Congress Office of Technology Assessment: Reviewer and panelist, Making the ADA work for people with psychiatric disabilities in the workplace, 1993.

Workshop presenter in areas of testing, employment selection and litigation, testing and public policy, ADA, work-based learning, testing standards, SIOP Principles, diversity in admissions, ethics in assessment, predictive validity, admissions testing, higher educational assessment, and research funding at regional applied I-O meetings and conferences.

ELECTED POSITIONS:

American Educational Research Association, Division D (Measurement and Research Methodology), Vice President and Council Representative, 2012-2015.

American Psychological Association: Council of Representatives, 1997-2003 (SIOP).

Association of Test Publishers: Board of Directors 2004 – 2008; Chair 2007; Treasurer 2008-2010.

Division of Evaluation, Measurement, and Statistics, American Psychological Association: President, 2000-01, Member-at-Large, 1997-2000.

Division of General Psychology, American Psychological Association: Member-at-Large, 2002-2005.

National Council on Education in Measurement: Board of Directors, 2002 – 2005, 2009-2012. President 2010-11.

Society for Industrial and Organizational Psychology: Member of Executive Committee, 1988- 2003; Council Representative, 1997-2003; Member-at-Large, 1995-97

TESTIMONY:

California Legislature on Test validity and consequences of subgroup differences in ability testing, 1997.

Invited testimony before the National Commission on Testing and Public Policy, 1989

Maine Joint Committee on Education and Cultural Affairs on the subject of Legislative Document No. 843 -
- H.P. 1283, January 17, 2006.

Michigan Senate Education Committee, on the replacement of the MAEP and the use of admissions tests
for accountability, April 22, 2004.

National Advisory Commission on Work-Based Learning, 1992 – 93.

National Assessment Governing Body, panel on testing persons with disabling conditions, October 14,
1998.

National Research Council's Committee on National Research Service Awards, May 1993.

Nevada Legislative Hearing on College and Career Readiness, Reno, NV., May 2012.

New York Assembly Committee, Test Disclosure, 1990.

New York Senate Committee, Proposed legislation to regulate admissions testing, 2006.

U.S. Congress, House Education and Labor Subcommittee, Goals 2000, 1993.

U.S. Congress, House Appropriations Subcommittee, Research Funding in Behavioral Sciences, 1992. 1993.

U.S. Congress, Senate and House Committees, Prepared testimony for APA presented on Civil Rights Act,
Polygraph Protection Act, Integrity Testing, American 2000, Americans with Disabilities, and
Appropriations.

U.S. Department of Education Hearings on Common Assessments for College and Career Readiness,
November 2009.

EXTERNAL GRANTS (PROJECT DIRECTOR):

National Assessment Governing Board (NAGB) (2008-10). Co-Project Director. Alignment and linkage of
Twelfth grade NAEP and the SAT.

Southern Regional Educational Board (2000-01). Project Manager. Design and development of common
Algebra assessment and item bank.

Office of Educational Research and Improvement (1996-98). Project Manager. Research grant to examine
the generalizability and utility of local models for scoring performance assessments. Working
collaboratively with six school districts examining different models for local scoring of Pacesetter
culminating assessments.

Maryland Department of Education (1996-97). Project Director. Contract to design Maryland's High School
Assessment System. Designing requirements and specifications for an end-of-course assessment system
for high school graduation and higher education uses. Conducting public engagement with stakeholder
groups and advising the state board.

National Institute of Occupational Health and Safety (1992-1994). Project Director. Cooperative agreement to develop a model interdisciplinary program to train doctoral level psychologists in occupational health psychology and disseminate research on preventive interventions to policymakers, psychologists and researchers.

Department of Labor, (July, 1992-93). Project Director. Grant to support a review of methodologies and strategies in cognitive psychology and job analysis appropriate for the next revision of the "Dictionary of Occupational Titles."

National Institute on Drug Abuse, (February 1992). Co-Project Director. Examination of awareness and knowledge of the mechanisms for receiving outside funding to support research by recent doctoral degree recipients in psychology.

National Science Foundation, Principle Investigator or Co-PI on several contracts related to AP Redesign and Instructional development.

PRODUCT DEVELOPMENT:

Business and Project Plan for Computerized CLEP Examinations. Award for New Product Development, Educational Testing Service, 1998.

Led CB/ETS psychometric/research and redesign teams for the 2005 SAT with writing.

Prototype of Non-cognitive assessments for college admissions. Pilot testing in 2007-08 with applicants across 13 colleges.

Psychometric Research and Design of AP Potential Software. Product introduced by College Board in 2001 for expanding access in AP Courses and Examinations based on prior accomplishments and test performance.

Study Skills Inventory for high school and college freshmen. Prototype completed and product in development, 2005-2009.

SELECTED BIBLIOGRAPHY:

Camara, W.J., O'Connor, R., Mattern, K., and Hanson, M.A. (Eds.). (2015). Beyond academics: A holistic framework for enhancing education and workplace success. ACT Research Report 2015 (4). Iowa City, IA: ACT. Retrieved from http://www.act.org/research/researchers/reports/pdf/ACT_RR2015-4.pdf

Mattern, K., Burrus, J., Camara, W.J., O'Connor, R., Hanson, M.A., Gambrell, J., Casillas, A., and Bobek, B. (2014). Broadening the definition of college and career readiness: A holistic approach. ACT Research Report 2014 (6). Iowa City, IA: ACT. Retrieved from http://www.act.org/research/researchers/reports/pdf/ACT_RR2014-5.pdf

[R]Camara, W.J. (2014). Issues facing testing organizations in using the Standards for Educational and Psychological Testing. *Educational Measurement: Issues and Practice*, 33 (4) 13-15.

[R] Camara, W.J. (2013). Defining and measuring college and career readiness: A validation framework for new State consortium assessments. *Educational Measurement: Issues and Practice*, 32 (4) 16-27.

[R] Camara, W.J., Packman, S. and Wiley A. (2013). College, graduate and professional school admissions testing. In K. Geisinger (Ed.), *Handbook of Testing and Assessment in Psychology* (pp. 297-318). Washington, D.C: American Psychological Association.

[R] Camara, W.J. and Shaw, E. (2012). Tests, score reports, research and getting along with the media. *Educational Measurement: Issues and Practice*.

[R] Harris, W.G., Jones, J.W., Klion, R., Arnold, D.W., Camara, W.J., and Cunningham, M. R. (2012). Test publisher's perspective on "An updated meta-analysis" of integrity testing. *Journal of Applied Psychology*, 97 (3), 531-536.

[R] Mattern, K., Kobrin, J., and Camara, W.J. (2012). Promoting Rigorous Validation Practice: An Applied Perspective. *Measurement: Interdisciplinary Research and Practice*, 10, pp 88-92.

Camara, W.J. and Quenemoen, R. (2012). *Defining and Measuring College and Career Readiness and Informing the Development of Performance Level Descriptors (PLDs)*. Commissioned white paper for PARCC. Available at <http://www.parcconline.org/sites/parcc/files/PARCC%20CCR%20paper%20v14%201-8-12.pdf>

Wyatt, J., Wiley, A., Camara, W.J., and Proestler, N. (2011). *The development of an index of academic rigor for college readiness*. College Board Research Report (2011-11). New York, NY: College Board. Available at <http://professionals.collegeboard.com/profdownload/pdf/RR2011-11.pdf>

Luecht, R. and Camara, W.J. (2011) *Evidence and design implications required to support comparability claims*. Commissioned white paper for PARCC. Available at http://parconline.org/sites/parcc/files/PARCC_WhitePaper-RLuechtWCamara%5B5%5D.pdf

Wyatt, J., Kobrin, J., Wiley, A., Camara, W.J., and Proestler, N. (2011). *SAT Benchmarks: Development of a college readiness benchmark and its relationship to school performance*. College Board Research Report (2011-5). New York, NY: College Board. Available at <http://professionals.collegeboard.com/profdownload/pdf/RR2011-5.pdf>

Wiley, A., Wyatt, J., and Camara, W.J. (2010). *Development of a multidimensional index of college readiness*. College Board Research Report (2010-03). New York, NY: College Board. Available at http://professionals.collegeboard.com/profdownload/pdf/10b_3110_CollegeReadiness_RR_WEB_110315.pdf

[R] Packman, S.J. Camara, W.J. and Huff, K., (2010). A snapshot of industry and academic compensation in educational measurement and assessment. *Educational Measurement: Issues and Practice*, Fall.

[R] Camara, W. J. (2009). Validity Evidence in Accommodations for English Language Learners and Students Disabilities. *Journal of Applied Testing Technology*, 10 (2). <http://www.testpublishers.org/jattmain.htm>

Mattern, K. Kobrin, J., Patterson, B., Shaw, K. and Camara, W.J. (2009). Validity is in the eye of the beholder: Conveying SAT research findings to the general public. In R. Lissitz (ed.) *The Concept of Validity: Revisions, New Directions and Applications*. Charlotte, NC: Information Age Publishing

Camara, W.J. (2009). College Admission Testing: Myths and Realities in an Age of Admissions Hype. In R. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 147-180). Washington, DC: American Psychological Association.

[R]Camara, W. J. and Lane, S. (2006). A historical perspective and current views on the Standards for Educational and Psychological Testing. *Educational Measurement: Issues and Practice*, 25, pp. 35-41.

Camara, W.J. (2006). Improving Test Development, Use, and Research: Psychologists in Educational and Psychological Testing Organizations. In R. Sternberg (Ed.), *Careers in Psychology*. Washington, DC: American Psychological Association.

[R]Phillips, S. & Camara, W. J. (2006). Legal and ethical issues in testing. In R. Brennan (Ed.), *Educational Measurement* (Volume IV) (pp. 733-755). AERA and American Council on Education.

Camara, W. J. & Kimmel, E. (Eds.) (2005). *New tools for admissions to higher education*. Mahwah, NJ: Erlbaum.

Camara, W. J. (2005). Broadening criteria of college success and the impact of cognitive predictors in admissions testing (pp. 81-107), In W.J. Camara & E. Kimmel (Eds.), *New tools for admissions to higher education*. Mahwah, NJ: Erlbaum.

Camara, W. J. (2005). Broadening predictors of college success (pp. 53-80). In W.J. Camara & E. Kimmel (Eds.), *New tools for admissions to higher education*. Mahwah, NJ: Erlbaum.

Korbrin, J., Camara, W.J., & Milewski, G. (2004). The utility of the SAT I and SAT II for admissions. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 251-276). New York: Routledge Falmer.

Schmidt, A.E., & Camara, W.J. (2004). Group differences in standardized test scores and other educational indicators. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 189-202). New York: Routledge Falmer.

[R]Cahalan, C., Mandinach, E. & Camara, W.J. (2003). The impact of flagging on the admissions process. *Journal of College Admissions*. No. 186.

Camara, W.J. (2003). What educators need to know about professional testing standards. In J. Wall & G. Walz (Eds.), *Measuring up: Resources on testing for teachers, counselors, and administrators*. Greensboro, NC: ERIC/CASS.

Noble, J. and Camara, W. (2003). Issues in college admissions testing. In J. Wall & G. Walz (Eds.), *Measuring up: Resources on testing for teachers, counselors, and administrators*. Greensboro, NC: ERIC/CASS.

Camara, W.J., Kimmel, E., Scheuneman, J., and Sawtell, E. (2003). Who's grades are inflated? *College Board Research Report (2003-4)*. New York: College Board. Available at http://professionals.collegeboard.com/profdownload/pdf/04843cbreport20034_31757.pdf

Camara, W.J. (2003). Construct validity. In R. F. Ballesteros (Ed.) *Encyclopedia of Psychological Assessment* (Volume 2) (pp. 1070-1075). London: Sage Publications.

Camara, W.J. (2002). Advances in scoring and inferences concerning examine behavior in computer-adaptive testing. In Mills, C., Potenza, M, and Ward, W. (Eds.) *Computer-Adaptive Testing: Building a foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Linn, R.L., Drasgow, F., Camara, W., Crocker, L., Hambleton, R.K., Plake, B.S., Stout, W. and van der Linden, W.J. (2002). Examinee behavior and scoring computer-based tests. In C. Mills, M. Potenza, J. Fremer and W. Ward (Eds.) *Computer-based testing: Building the foundation for future assessment*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Noble, J., Camara, W., and Fremer, J. (2002). Admissions testing and students with disabilities. In Ekstrom, R., and Smith, D. (Eds.) *Assessing individuals with disabilities* (pp. 173-190). Washington, DC: American Psychological Association.

Mandinach, E.B., Cahalan, C. and Camara, W.J. (2002). The impact of flagging on the admissions process: Policies, practices and implications. *College Board Research Report (No. 02-02)*. New York: College Board. Available at <http://www.ets.org/Media/Research/pdf/RR-05-20.pdf>

Cahalan, C., Mandinach, E.B., and Camara, W.J. (2002). Predictive validity of STA I: Reasoning test for test takers with learning disabilities and extended time accommodations. *ETS Research Report (No. 02-03)*. Princeton, NJ: ETS. Available at <http://www.ets.org/Media/Research/pdf/RR-02-03-Mandinach.pdf>

[R]Scheuneman, J.D., Camara, W.J., Cascallar, A.S., Wendler, C., and Lawrence, I (2002). Calculator access, use and type in relation to performance on the SAT I: Reasoning test in mathematics. *Applied Measurement in Education*, 15 (1), 95-112.

Camara, W.J. and Echternacht (2000). The SAT I and high school grades: utility in predicting success in college. *College Board Research Note (RN-10)*. New York: College Board.

Camara, W. (2000). Using class rank alternative plans for college admissions. *Association of American Colleges and Universities: Diversity Digest, Summer*, 8-10.

[R]Camara, W.J., Dorans, N., Morgan, R. and Myford, C (2000). Advance Placement: Access not quality. *Educational Policy Analysis Archives*. 8 (40). Online journal, available: <http://epaa.asu.edu/epaa/v8n40.html>

[R]Camara, W.J. and Merenda, P.M. (2000). Using personality tests in preemployment screening: Issues raised in "Soroka v. Dayton-Hudson Corporation." *Psychology, Public Policy and the Law*, 6, (4), 1-23.

[R]Camara, W., Puente, A., and Nathan, J. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice*, 31 (2), 141-154.

Camara, W. and Schmidt, A. (1999). Group differences in standardized testing and social stratification. *College Board Research Report (No. 99-5)*. New York: College Board.

[R]Schneider, D.L., Camara, W.J., Tetrick, L. and Sternberg, C. (1999). Training in occupational health psychology: Initial efforts and alternative models. *Professional Psychology: Research and Practice*, 30 (2), 138-142.

Nathan, J., and Camara, W.J. (1999). Concordance of the examinee performance on the SAT and ACT. *College Board Research Note (RN99-7)*. New York: College Board.

Powers, D. and Camara, W. (1999). Coaching and the SAT I. *College Board Research Summary (RN99-6)*. New York: College Board.

Camara, W.J. and Millsap, R. (1998). Using the PSAT/NMSQT and course grades in predicting success in Advanced Placement. *College Board Research Report* (RR98-5). New York: College Board.

Camara, W.J., Copeland, T. and Rothschild, B. (1998). Effects of extended time on the SAT I: Reasoning Test score growth for students with learning disabilities. *College Board Research Report* (No. 98-7). New York: College Board.

Nathan, J., and Camara, W.J. (1998). Score change when retaking the SAT I Reasoning Test. *College Board Research Note* (RN98-5). New York: College Board.

Camara, W. (1998). High school grading policies. *College Board Research Note* (RN98-4). New York: College Board.

Smith, R. and Camara, W. (1998). Block schedules and student performance on AP examinations. *College Board Research Note* (RN98-3). New York: College Board.

Camara, W., Kimmel, E., et. al., (1997). Design of a high school assessment system. (Vols. I and II). *Technical Report of the College Board and ETS*. Baltimore, MD. Maryland State Department of Education.

[R]Camara, W. J. (1997). Educational assessment: Responsible uses and professional dilemmas. *European Journal of Psychological Assessment*, 13 (2), 140-152.

Camara, W.J. and Kraiger, K. (1997). Organisational infrastructure for selection and assessment in the USA. In Smith, M & Sutherland, V. (Eds.). *Professional issues in selection and assessment* (pp. 139-146). Wiley: London.

Camara, W. (1997). Validity, Fairness, and Public Policy of Employment Testing: Influences of the American Psychological Association (pp. 3-11). In Barrett, R., *Fair employment strategies*.

Camara, W., Kimmel, E. and colleagues (1997). *Models for the Design of Maryland's High School Assessments*. College Board Technical Report (CBTR97-1).

[R]Camara, W.J. and Schneider, D. (1995). Questions of construct breadth and openness of research on integrity tests. *American Psychologist*, 47 (3).

[R]Camara, W. J. and Brown, D. (1995). Educational and employment testing: Changing concepts in measurement and policy. *Educational Measurement: Issues and Practice*, 14 (1), 5-12.

Camara, W. J. (1995). APA involvement in employment testing policy and litigation: An historical overview. Unpublished manuscript.

[R]Camara, W.J. and Schneider, D. (1994). What we know and still don't know about integrity tests. *American Psychologist*, 47 (3).

Camara, W.J., and Baum, C. (1993). *Developing careers in research: Knowledge, attitudes and intentions of recent doctoral recipients in psychology*. (Final report 92MF04400101D) Rockville, MD, National Institute of Drug Abuse.

Camara, W.J. (1992). Fairness and "fair-use" in employment testing: A matter of perspectives (pp. 215-233). In Geisinger, K, *Testing of Hispanics*. Washington, DC: APA

[R]Camara, W.J. (1991). A national exam: Has its time come? *Child Behavior & Development*, 7 (9-10).

[R]Camara, W.J., et. al. (1990). Enhancing psychological science: A report by the Science Advisory Committee. *American Psychologist*, 45 (7).

[R]Fremer, J., Diamond, E., and Camara, W. (1989). Developing a "Code of Fair Testing in Education." *American Psychologist*, 44 (7), 1062-1067.

[R]Bond, L., Camara, W.J., and VandenBos, G.R. (1989). Psychological test standards and clinical practice. *Hospital and Community Psychiatry*, 40 (7), 687-693.

Camara, W.J. (1989). *Detecting dishonest employees: What is the state of the art?* Proceedings of the Second Annual National Assessment Conference, (pp.26-28) University of Minnesota and Personnel Decisions Inc., Minneapolis, MN.

Camara, W.J., Kuhn, D., and Ziemak, J. (1987). Development and training of Army civilian first-line supervisors. (Final Report FR-87-36). Alexandria, VA. Human Resources Research Organization.

[R]Waters, B.K., Laurence, J.H., & Camara, W.J. (1987). *Personnel enlistment and classification procedures in the U.S. Military*. Washington, D.C.: National Academy of Science Press.

Camara, W.J., & Laurence, J.H. (1987). Military classification and high aptitude recruits (Technical Report TR-PRD-87-16). Alexandria, VA. Human Resources Research Organization.

Camara, W.J. (1986). The effects of job previews on self-selection decisions. *Dissertation Abstracts International*, 47, DA8623268.

Camara, W.J. (1984). Assessment centers: A critical review of the literature. Unpublished Paper, Champaign-Urbana: University of Illinois.

Camara, W.J. (1983). Personnel selection: A classification and review of techniques. Unpublished Masters Thesis, Champaign-Urbana: University of Illinois.

Camara, W.J. (1981) Infusion - inservice: Career awareness. A Massachusetts guide: Promising practices in career education. Boston, MA: Department of Education.

SELECTED PRESENTATIONS:

Camara, W.J. (2015). Employing empirical data in judgmental processes. Paper presented at the National Conference on Student Assessment, San Diego, CA.

Camara, W.J. and Westrick, P. (2015). Admissions testing in the United States. Invited presentation at the Annual Meeting of the American Educational Research Association in Chicago, IL.

Camara, W.J. (2015). "Evidentiary basis related to claims concerning college and career readiness." Colloquium, University of Massachusetts, Amherst, Graduate Program in Education.

Camara, W.J. (2015). Overview of the 2014 Revision of the 'Standards for Educational and Psychological Testing.' Paper presented at the Association of Test Publishers, Palm Springs, CA.

Camara, W.J. (2014). Test security: Prevention-detection-investigation. Workshop presentation for the Minnesota State Department of Education, Offices of Assessment and Accountability.

Camara, W.J. (2014). How has our approach to test security evolved and where are we headed. Paper presented at the Conference on Test Security, Iowa City, IA.

Camara, W.J. (2014). Developing sources of validation evidence across assessment settings. Invited presentation at the International Testing Commission, San Sebastian, Spain.

Camara, W.J. and Shaw, D. (2014). Use of comment codes during performance scoring to provide formative feedback. Paper presented the National Conference on Student Assessment, New Orleans, LO.

Camara, W.J. (2014). Employing empirical data in judgmental standard setting processes. Paper presented at the Annual Meeting of the Society for Industrial and Organizational Psychology, Honolulu, HI.

Camara, W.J. (2014). Fisher v. University of Texas: The future of affirmative action. Participant in panel at the Annual Meeting of the Society for Industrial and Organizational Psychology, Honolulu, HI.

Camara, W.J. (2014). AERA Vice-Presidential Symposia: Technology Enhanced Items in Large Scale Assessments. Annual Meeting of the American Educational Research Association, Philadelphia, PA.

Camara, W.J. (2013). PISA's use for international benchmarking and comparisons of post-secondary readiness. Invited panel, Oxford University.

Camara, W.J. (2013). College and career readiness: Criterion-related outcomes. Invited address at the Maryland Assessment Research Center for Educational Success, University of Maryland at College Park.

Camara, W.J. (2013). Implications of consortia assessments for Higher Education. Paper presented at the National Conference on Student Assessment at the National Harbor, MD.

Camara, W.J. (2013). Innovations in psychometrics and assessment. Developing college and career readiness assessments. Workshop at the National Center for Measurement in Education, San Francisco, CA.

Camara, W.J. (2012). Admissions practices and college going in the U.S. Invited presenter at the International Conference on Assessment and Evaluation, Riyadh, Saudi Arabia. National Center on Assessment in Higher Education.

Reshetar, R., and Camara, W. J. (2012). Redesigning the Advanced Placement Science Assessments application of evidence centered design. Invited Panelist at the National Research Council Workshop.

Camara, W.J. (2012). College and career readiness: Establishing validation evidence to support the use of new assessments. Invited lecture at the Pearson Center for Applied Psychometric Research, University of Texas at Austin.

Camara, W.J. (2012). Defining and measuring college and career readiness: Developing Performance level descriptors and defining criteria. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, Canada.

Camara, W.J. (2012). Invited panel presentation on data integrity and cheating. National Center on Educational Statistics. Sponsored Symposium on Testing Integrity, Washington, D.C.

Camara, W.J. (2011). College and career readiness: An initial validation argument. Paper presented at the National Conference on Student Assessment, Orlando, FL.

Camara, W.J. (2011). Developing and expanding state K-20 longitudinal data systems: Common core state standards and consortia assessments. Paper presented at the National Conference on Student Assessment, Orlando, FL.

Camara, W.J. (2011). The revised testing standards: Potential impact and consequences for assessments in employment and business settings. Invited address at the International Personnel Assessment Council, Washington, DC.

Camara, W.J. (2011). College and career readiness standards and assessments: An initial validation argument. Paper presented at the CCSSO National Conference on Student Assessment, Orlando, FL.

Camara, W. J. (2011). Empirical benchmarks in a judgmental standard setting process. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Camara, W. J. (2011). Uncovering Educational Measurement & Assessment Professionals: Demographics, Education, Experience and Engagement. Presidential Address at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Camara, W.J. (2011). Formative assessment: Implications of the common core on classroom assessment. Invited address at the Annual Meeting of the American Educational Research Association, Classroom Assessment SIG.

Camara, W. J., Wiley, A., Wyatt, J., and Kobrin, J. (2011). College readiness benchmarks. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Camara, W.J. (2010). Validating claims and evidence related to student college and career readiness: Lessons learned from higher education. Invited presentation at the Annual CCSSO Policy Meeting, Louisville, KY.

Camara, W.J. (2010). Developing benchmarks for college and career readiness. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Camara, W.J. (2010). Multidimensional models of college readiness. Paper presented at the Large Scale Assessment Conference, Detroit, MI.

Camara, W. J. (2010). Progress in revising the Standards for Educational and Psychological Testing. the Annual Conference of the American Psychological Association, San Diego, CA.

Camara, W.J. (2009). Operational Issues in Developing National Admissions Testing and College Credit Testing Programs in the U.S. Invited Colloquium at the University of Aachen, Germany.

Camara, W.J. (2009). Common Core Standards and Coordinated State Assessment. Invited Symposium at the Annual Meeting of the American Educational Research Council, Denver, CO.

Camara, W.J. (2009). You can get there from here: Innovation in Educational assessment and linking accountability tests. Invited address at the National Conference of State Legislators, Washington, DC.

Camara, W. J. (2009). Noncognitive assessments in college admissions. Paper presented at the Annual Conference of the American Psychological Association, Toronto, Canada.

Camara, W., Kobrin, J., Mattern, K., Patterson, B., and Shaw, E. (2008). The Long and Winding Road: Researching the Validity of the SAT. Invited paper at the 9th annual conference of the Maryland Assessment Research Center for Education Success (MARCES), College Park, MD.

Camara, W. J. (2008). Innovations in assessment. Presenter at the Invitational Conference Educational Testing in America: State Assessment, Achievement Gaps, Federal Policy and Innovations, Sponsored by ETS and the College Board, Washington, DC.

Camara, W. J. (2008). College readiness vs college admissions: Will we ever resolve the chasm between the K-12 and Higher Education? Invited address at Invitational Conference on Defining Enrollment in the 21st Century, sponsored by the University of Southern California's Center for Enrollment Research, Policy and Practice.

Camara, W. J. (2008). Diversity in admissions. Invited Address at the ETS Conference of Institutional Researchers, Measuring Success and Making Assessment Data Work at Your Institution, Princeton, NJ.

Camara, W. J. (2008). The educational measurement profession: state of our art. Presentation at the annual meeting of the National Council on Measurement in Education, New York, NY.

Camara, W. J. (2007). Protecting test takers. Invited Presidential Symposium at the Annual Conference of the American Psychological Association, San Francisco, CA.

Camara, W. J. (2007). Revising the standards for educational assessment. Invited Symposium at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Camara, W. J. (2006). Using norm referenced tests for accountability under NCLB. Presenter at the Annual Meeting of the National Association of Collegiate Admissions Counselors, Pittsburgh, PA.

Camara, W. and Schmidt, A. (2006). University Admissions Practices in the US and the Role of Admissions Tests. Invited Address at UCAS Conference, Nottingham, United Kingdom.

Camara, W J. (2005). Constraints in current admissions practices: Impacts on diversity and definition of college success. Invited Address at the Goldman-Sachs Foundation and ETS Symposium on Addressing Achievement Gaps, Princeton, NJ.

Camara, W. J. (2005). Update on the new SAT. Annual Meeting of the National Association of Collegiate Admissions Counselors, Tampa, FL.

Camara, W. J. (2005). Design and development of the SAT Writing Test. National Council on Measurement in Education, Montreal, Canada.

Camara, W.J., Kobrin, J., and Sathy, J (2005). Is there an SES advantage on the SAT and college performance? National Council on Measurement in Education, Montreal, Canada.

Camara, W. (2004). The Use of Qualitative and Quantitative Data in Admissions. Annual Meeting of the National Association of Collegiate Admissions Counselors, Milwaukee, WI.

Laitusus, V. Camara, W. J. and Wang, B. (2004). An examination of differential item functioning for language minorities on a verbal and math reasoning test. National Council on Measurement in Education, San Diego, CA.

Camara, W.J. (2004). New predictors in college admissions. Annual Meeting of the Society of Industrial and Organizational Psychologists, Chicago, IL.

Camara, W. J. (2003) Current tests and future designs in admissions testing: The new SAT. CASMA-ACT Invitational Conference. Iowa City, IA.

Camara, W. J. (2003). Validity and utility of admissions tests. Invited symposium, American Council on Education, Washington, DC.

Camara, W.J. (2003). Changes to the SAT: National Association of Collegiate Admissions Counselors.

Camara, W. (2003). Making test results more useful and understandable: Advances in diagnostic score reporting. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Camara, W. (2002). Revision of the Principles for the Validation and Use of Personnel Selection Procedures, Workshop conducted at the Mid-Atlantic Personnel Assessment Consortium, New York, NY.

Camara, W. (2002). Predicting success in employment and education: Uses and limitations of tests and other factors. Invited address, New York Academy of Sciences.

Camara, W. (2002). Prediction and testing. Invited address at the CRESST Conference on Assessment, Accountability and Improvement, Los Angeles, CA.

Camara, W. (2002). Admissions tests : Use and value in higher education. Invited address the Association of American Universities, Meeting of Presidents and Chancellors, Atlanta, GA.

Camara, W. (2002). The future of admissions testing. Annual Meeting of the National Association of Collegiate Admissions Counselors, Salt Lake City.

Camara, W. (2002). Fairness in employment testing. Paper presented at the Annual Convention of the American Psychological Association, Chicago, IL.

Camara, W. (2002). Testing and admissions in higher education. Invited presentation at the Annual Meeting of the American Association for the Advancement of Science, Boston, MA.

Camara, W. (2001). The utility of the SAT I and SAT II for admission at the University of California and the nation. Paper presented at the Invitational Conference on Rethinking the SAT in university admissions, University of California at Santa Barbara.

Camara, W. (2001). Test preparation on the SAT: Impact on validity. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Camara, W. (2001). Utility of the SAT in college admissions. Colloquium at the University of California at Davis.

Camara, W. (2001). Do accommodations improve or hinder psychometric qualities of assessment? Presidential address for Division 5 at the Annual Convention of the American Psychological Association, San Francisco, CA.

Camara, W. (2000). Future of educational assessment. Paper presented at the Annual Convention of the American Psychological Association, Washington, D.C.

Camara, W. (2000). Implications of the revised testing standards for personnel selection. Invited Address at the Annual Conference of the International Personnel Management Assessment Council, Washington.

Camara, W. (2000). Performance of test takers with LD or ADD on the SAT and subsequent college behavior. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Camara, W. (2000). Implications of the testing standards in personnel assessment. Presentation at the Annual Meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Camara, W. (1999). The revised 'Standards for Educational and Psychological Testing'. Workshop at the Mid-Atlantic Personnel Assessment Consortium, New York, NY.

Camara, W. (1999). Retesting on the SAT under standard and non-standard administrations. Presentation at the National Conference of Measurement in Education, Montreal, Canada.

Camara, W. (1999). Testing practices in clinical assessment. Paper presented at the American Psychological Association, Boston, MA.

Camara, W. (1998). Accommodations for persons with disabilities: results of attempts to establish comparability in cognitive testing. Continuing education workshop at Personnel Testing Council of Metropolitan Washington.

Camara, W. (1998). Alternatives to item pattern scoring and use of response-time estimation in computer adaptive testing: Invited Presentation. ETS Invitation Conference on future assessments, Philadelphia, PA.

Camara, W. (1998). Future trends in assessment. Presentation at the Annual Conference for the Society for Industrial and Organizational Psychology, Dallas, TX..

Camara, W. (1998). Selection into I-O programs: Focus on GRE validity. Symposium at the Annual Conference for the Society for Industrial and Organizational Psychology, Dallas, TX

Camara, W. (1998). Rights and responsibilities of test takers. Presentation at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Scheuneman, J. and Camara, W. (1998). Analysis of mathematics achievement in Pacesetter program. Presentation at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Camara, W. (1998). Evaluating math curricular reform efforts. Presentation at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Camara, W. (1998). Psychometric and operational constraints remaining in CBT. Colloquium at Fordham University Graduate Departments of Psychology and Education, New York, NY.

Camara, W. (1997). State and district accountability: Uses and misuses of assessments. Presentation at the Large Scale Assessment Conference of the Council of Chief State School Superintendents, Colorado Springs, CO.

Camara, W. (1997). Effects of calculator use on performance of on a mathematics admissions test. Panel discussant at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Camara, W. (1997). Assessing workplace skills: Public policy and technical considerations. Colloquium at Baruch College, City University of New York.

Camara, W. (1996). Effects of extended time on the performance of students with disabilities. Paper presented at the Annual Conference of the National Association of College Admissions Counselors, Minneapolis, MN.

Camara, W. (1996). Flagging test scores for students with disabilities: Understanding and using test scores for admissions and placement decisions. College Board National Forum, New York, NY

Camara, W. (1996). Adapting/ Translating educational and psychological tests: Issues, technical advances, and guidelines. Panel discussion at the Annual Convention of the American Psychological Association, Toronto, Canada.

Camara, W. (1996). Doctoral training in organizations: Comparisons among Business schools and Psychology departments. Panel discussion at the Annual Meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Camara, W. (1996). SCANS based competencies: Discussion of the national job analysis project. Discussant at the Large Scale Assessment Conference of the Council of Chief State School Superintendents, Phoenix, AZ.

Camara, W. (1995). Test speededness and other implications of testing persons with disabilities in large-scale programs. Paper presented at the October Meeting of the Personnel Testing Council, Washington, DC., and (1996) Mid-Atlantic Personnel Assessment Consortium, Potomac, MD.

Camara, W. (1995). Flagging of test scores: Policies, Data and Opportunities. Panel discussion, ETS Committee for People with Disabilities, Educational Testing Service, Princeton, NJ.

Camara, W. (1995). Lessons for test developers from the NACAC Commission on Standardized Testing. Paper presented at the Annual Conference of the National Association of College Admissions Counselors, Boston, MA.

Camara, W. (1995). Standard setting: A mixed bag of judgment, psychometrics and policy. Paper presented at the Annual Convention of the American Psychological Association, New York, NY.

Camara, W. (1995). Federal funding opportunities in Industrial and Organizational Psychology (moderator and presenter). Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Camara, W. (1995). The New SAT: Reactions (chair) Symposium at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Camara, W. (1994). International perspectives on test use: Options in education and enforcement? Presentation at the 23rd International Congress of Applied Psychology, Madrid Spain.

Camara, W. (1994). Developments in creating the new national database of occupational titles (chair and presenter). Society for Industrial and Organizational Psychology, Nashville.

Camara, W. (1994). The impact of national testing standards on personnel assessment. Invited presentation at the International Personnel Management Association Assessment Council Meeting, Charleston, SC.

Camara, W. (1994). Test standards: Balancing technical, applied and policy issues. Invited presentation, Personnel Testing Council of Washington, DC.

Camara, W. (1993). Who should control access and use of Neuropsychological Tests? Invited presentation at the National Academy of Neuropsychology, Phoenix, AZ.

Camara, W. (1993). Implications of the "Americans with Disabilities Act" on Assessment, Invited address at International Personnel and Management Association, Sacramento, CA.

Camara, W. (1993). Ethical issues in research, teaching, and publication for industrial psychologists (chair, panelist). Society for Industrial and Organizational Psychology, San Francisco, CA.

Camara, W. (1993). I/O Psychology in the Public-Policy-Making Process (panelist). Society for Industrial and Organizational Psychology, San Francisco, CA.

Lipsitt, L. & Camara, W.J. (1993). The childhood origins of creativity. Paper presented at the Nebraska symposium on gifted children, Lawrence, Kansas.

Camara, W. (1992). 100 Years of Psychological Testing (chair). Centennial Convention of the American Psychological Association, Washington, DC.

Camara, W. (1992). Occupational Health Psychology: A new specialty for psychology and training needs. Centennial Convention of the American Psychological Association, Washington, DC.

Camara, W. (1992). Correlates between personnel and educational assessment in national policy. Invited Address at the Annual Convention of the American Psychological Society, San Diego, CA.

Camara, W.J. (1992). Affirmative Action and the Civil Rights Act of 1991. Invited address at International Personnel and Management Association, Baltimore, MD.

Camara, W.J. (1992). Americans with Disabilities Act and the Civil Rights Act of 1991: Implications for industrial psychologists. Annual Conference of the Society of Industrial and Organizational Psychology, Montreal, Quebec.

Camara, W.J. (1992). Scans and America 2000. Annual Conference of the Society of Industrial and Organizational Psychology, Montreal, Quebec.

Camara, W.J. (1991). Federal funding opportunities at ADAMHA, the Department of Energy and the Department of Agriculture (moderator). Thirty-sixty Institute on Federal Funding, National Graduate University, Washington, D.C.

Camara, W.J. (1990). Disclosure of test scores, items and protocols in educational settings. Paper presented at the 98th Annual Convention of the American Psychological Association, Boston, MA.

Camara, W.J. (1991). Integrity testing: Risks and rewards. Invited address, Personnel Testing Council of Washington, DC.

Camara, W.J. (1989). Detecting dishonest employees: What is the state of the art? Paper presented at the Annual National Assessment Conference of the University of Minnesota and Personnel Decisions Inc., Minneapolis, MN.

Camara, W.J. (1989). Predicting Honesty: Scientific Evidence, Business Necessity, and Social Policy Issues. Paper presented at the 97th Annual Convention of the American Psychological Association, New Orleans, LO.

Camara, W.J. (1989). Legal burden in employment selection: Recent court decisions. Symposium at the 4th Annual Convention of the Society of Industrial/Organizational Psychologists, Boston, MA.

Camara, W.J. and Kuhn, D. (1988). Development of a mixed standard rating scale for training and development. Paper presented in Division 14 of the 96th Annual Convention of the American Psychological Association, Atlanta, GA.

Camara, W.J. (1987). The utility of a job-person match for personnel selection decisions. Paper presented in Division 14 of the 95th Annual Convention of the American Psychological Association, New York, NY. (ERIC Document Reproduction Service No. ED 289 149).

Ziemak, J.P., Camara, W.J., Fisher, G.P., and Darmsteadt, G.H. (1987). Development of effective army civilian first-line supervisors. Paper presented at the 29th Annual Military Testing Association Conference, Quebec, Canada.

Camara, W.J., Colot, P., Hutchinson, G., & Campbell, B. (1987). The reality of data collection. Paper presented at the 95th Annual Convention of the American Psychological Association, New York, NY. (ERIC Document Reproduction Service No. 290 775).

Camara, W.J. (1986). The utility of biodata in predicting military performance. Paper presented at the 26th Annual Military Testing Association Conference, Mystic, CT.

Camara, W.J. (1986). Effects of job previews on personnel selection. Paper presented at the National Conference of the Association of Human Resources Management and Organizational Behavior, New Orleans, LA.

Camara, W.J. (1986). Equivalence of rater sources on job analysis ratings. Paper presented in division 14 at the 94th Annual Convention of the American Psychological Association, Washington, D.C. (ERIC Doc.Reproduction Service No. ED 281 455).

Camara, W.J. and Means, B. (1986). Status of low-aptitude accessions following military service. Paper presented at the 94th Annual Convention of the American Psychological Association, Washington, D.C.

Additional presentations at national and regional meetings and university colloquium not normally cited.
9/2015

EXHIBIT MMM

Case No. 1:14-cv-00857-TSC-DAR

From: John S. Neikirk
To: 'Wayne Camara'
CC: Suzanne Lane (sl+@pitt.edu); dfrisbie@uiowa.edu; Felice Levine
Sent: 2/14/2014 9:40:11 PM
Subject: RE: Existing Standards

Hi Wayne, thanks for your message and letting me know that this is possible. We are in communication with AERA's legal counsel, and perhaps also some outside help as well. We will be in touch as soon as we learn more.

Best wishes, John

John Neikirk
 Director of Publications
 American Educational Research Association
 1430 K Street, NW, Suite 1200
 Washington, DC 20005
 202.238.3238
 jneikirk@aera.net

From: Wayne Camara [mailto:Wayne.Camara@act.org]
Sent: Friday, February 07, 2014 8:26 AM
To: John S. Neikirk
Cc: Suzanne Lane (sl+@pitt.edu); dfrisbie@uiowa.edu
Subject: RE: Existing Standards

John – the management committee has funds to cover any legal work that might be needed to get this issue escalated. So feel free to let us know if you want to bring in some legal counsel to pursue this.

From: John S. Neikirk [mailto:JNeikirk@aera.net]
Sent: Wednesday, February 05, 2014 10:35 AM
To: Felice Levine; Wayne Camara; Ernesto, Marianne; 'lwise@humrro.org'; SL@pitt.edu; Frisbie, David A; Gerald Sroufe
Cc: Barbara Plake
Subject: RE: Existing Standards

Hello everyone, I am following up on the messages from December about the pdf posting of Standards at:
https://law.resource.org/pub/us/cfr/ibr/001/aera_standards.1999.pdf

I had thought this was taken down after I wrote to the organization in December, asking that the pdf be taken down immediately. I checked yesterday and the pdf is still available. I wrote back to the organization, and I received a reply with the attached pdf letter, which had been sent in December apparently (the messages from this organization go straight into spam).

As you can see from the attached letter, the author (Carl Malamud of Public.Resource.Org) claims that the 1999 edition is in now in the public domain because it was incorporated by Reference by the Department of Education. Please see: <http://www.ecfr.gov/cgi-bin/text-idx?c=ecfr&sid=c67daaa426ddc5f9838ea4247c36c938&rgn=div8&view=text&node=34:3.1.3.1.34.10.39.8&idno=34>

We are looking into this now and will report back to the group when we have more information.

Best wishes, John

John Neikirk
 Director of Publications
 American Educational Research Association
 1430 K Street, NW, Suite 1200
 Washington, DC 20005



202.238.3238

ineikirk@aera.net

From: Felice Levine
Sent: Monday, December 16, 2013 2:38 PM
To: 'Wayne Camara'; Ernesto, Marianne; 'lwise@humrro.org'; SL@pitt.edu; Frisbie, David A; Gerald Sroufe
Cc: Barbara Plake; John S. Neikirk
Subject: RE: Existing Standards

Yes, wow.... This must be a copyright infringement or something to the equivalent.... But John will look into this, felice

From: Wayne Camara [<mailto:Wayne.Camara@act.org>]
Sent: Monday, December 16, 2013 2:16 PM
To: Ernesto, Marianne; 'lwise@humrro.org'; SL@pitt.edu; Frisbie, David A; Gerald Sroufe; Felice Levine
Cc: Barbara Plake
Subject: RE: Existing Standards

Jerry and Felice – can you look into this? We need to get them to pull this off the web.

From: Ernesto, Marianne [<mailto:MErnesto@apa.org>]
Sent: Monday, December 16, 2013 1:09 PM
To: 'lwise@humrro.org'; SL@pitt.edu; Wayne Camara; Frisbie, David A
Cc: Jerry Sroufe; Barbara Plake
Subject: RE: Existing Standards

This is news to me!

Jerry and all,
Any idea how this got posted?

Marianne

From: lwise@humrro.org [<mailto:lwise@humrro.org>]
Sent: Monday, December 16, 2013 1:53 PM
To: SL@pitt.edu; Wayne Camara; Frisbie, David A
Cc: Jerry Sroufe; Ernesto, Marianne; Barbara Plake
Subject: Existing Standards

Are they supposed to be available (in pdf form) free?

<https://law.resource.org/pub/us/cfr/ibr/001/aera.standards.1999.pdf>

Lauress (Laurie) L. Wise, Principal Scientist
Human Resources Research Organization (HumRRO)
20 Ragsdale Drive, Suite 260
Monterey, CA 93940
Phone: 831-647-1004
Fax: 831-375-4021
Cell: 703-727-3817

**UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA**

AMERICAN EDUCATIONAL RESEARCH)
ASSOCIATION, INC., AMERICAN)
PSYCHOLOGICAL ASSOCIATION, INC.,)
and NATIONAL COUNCIL ON)
MEASUREMENT IN EDUCATION, INC.,)
))
Plaintiffs,)
))
v.)
))
PUBLIC.RESOURCE.ORG, INC.,)
))
Defendant.)

Civil Action No. 1:14-cv-00857-TSC-DAR

**DECLARATION OF FELICE
J. LEVINE IN SUPPORT OF
PLAINTIFFS' MOTION FOR
SUMMARY JUDGMENT AND ENTRY
OF A PERMANENT INJUNCTION**

I, FELICE J. LEVINE, declare:

1. I am the Executive Director of the American Educational Research Association, Inc. ("AERA") I have been employed by the AERA since May 2002. I submit this Declaration in support of the motion of the AERA, the American Psychological Association, Inc. ("APA"), and the National Council on Measurement in Education, Inc. ("NCME") (collectively, "Plaintiffs" or "Sponsoring Organizations") for summary judgment and the entry of a permanent injunction.

2. As set forth in the AERA Bylaws, the Executive Director is the chief executive officer of the Association. In that capacity, I am responsible for all programmatic, financial, administrative, staffing, and managerial responsibilities of the AERA. I also advise on and implement the policies that guide our organization.

3. As publisher, the AERA has provided general oversight since November 1999 for the production, printing, sales, and marketing of the "Standards for Educational and Psychological Testing" (the "Standards"), and for the fiscal management of the revenue and expenditure of funds and resources of that publication. AERA was selected to serve as publisher

by the Management Committee of the three Sponsoring Organizations. As the Executive Director of the AERA, I have administrative oversight over all of AERA's implementation of its responsibilities regarding the Standards.

4. AERA is a District of Columbia not-for-profit corporation.

5. AERA is the major national scientific society for research on education and learning. AERA's mission is to advance knowledge about education, to encourage scholarly inquiry related to education, and to promote the use of research to improve education and serve the public good.

6. In 1955, Plaintiffs AERA and NCME prepared and published a companion document to APA's "Technical Recommendations for Psychological Tests and Diagnostic Techniques" (published in 1954), entitled, "Technical Recommendations for Achievement Tests." Subsequently, a joint committee of the three organizations modified, revised, and consolidated the two documents into the first Joint Standards. Beginning with the 1966 revision, the Sponsoring Organizations collaborated in developing the "Joint Standards" (or simply, the "Standards"). Each subsequent revision of the Standards has been careful to note that it is a revision and update of the prior version.

7. Beginning in the mid-1950s, the Sponsoring Organizations formed and periodically reconstituted a committee of highly trained and experienced experts in psychological and educational assessment, charged with the initial development of the Technical Recommendations and then each subsequent revision of the (renamed) Standards. These committees were formed by the Sponsoring Organizations' Presidents (or their designees), who would meet and jointly agree on the membership. Often a chair or co-chairs of these committees were selected by joint agreement. Beginning with the 1966 version of the Standards, this

committee became referred to as the “Joint Committee.”

8. Financial and operational oversight for the Standards’ revisions, promotion, distribution, and for the sale of the 1999 and 2014 Standards has been undertaken by a periodically reconstituted Management Committee, comprised of the designees of the three Sponsoring Organizations. As Publisher of the 1999 and 2014 Standards, AERA works in consultation with the Management Committee to implement its managerial guidance.

9. All members of the Joint Committee(s) and the Management Committee(s) are *unpaid* volunteers. The expenses associated with the ongoing development and publication of the Standards include travel and lodging expenses (for the Joint Committee and Management Committee members), support staff time, production, printing and shipment of bound volumes, and advertising costs. For the 2014 Standards, the production, printing and shipment of bound volumes, and advertising costs, are paid for by the publisher, AERA.

10. Many different fields of endeavor rely on assessments. The Sponsoring Organizations have ensured that the range of these fields of endeavor is represented in the Joint Committee’s membership – *e.g.*, admissions, achievement, clinical counseling, educational, licensing-credentialing, employment, policy, and program evaluation. Similarly, the Joint Committee’s members, who are *unpaid volunteers*, represent expertise across major functional assessment areas – *e.g.*, validity, equating, reliability, test development, scoring, reporting, interpretation, and large scale interpolation.

11. From the time of their initial creation to the present, the preparation of and periodic revisions to the Standards entail intensive labor and considerable cross-disciplinary expertise. Each time the Standards are revised, the Sponsoring Organizations select and arrange for extensive meetings of and work by the leading authorities in psychological and educational

assessments (known as the Joint Committee). During these meetings, certain Standards are combined, pared down, and/or augmented, others are deleted altogether, and some are created as whole new individual Standards. The 1999 version of the Standards is nearly 200 pages, took more than five years to complete.

12. The Standards were not created or updated to serve as a legally binding document, in response to an expressed governmental or regulatory need, nor in response to any legislative action or judicial decision. However, the Standards have been cited in judicial decisions related to the proper use and evidence for assessment, as well as by state and federal legislators. These citations in judicial decisions and during legislative deliberations occurred without any lobbying by the Plaintiffs.

13. AERA has not solicited any government agency to incorporate the Standards into the Code of Federal Regulations or other rules of Federal or State agencies.

14. Plaintiffs promote and sell copies of the Standards via referrals to the AERA website, at annual meetings, in public offerings to students, and to educational institution faculty. Advertisements promoting the Standards have appeared in meeting brochures, in scholarly journals, and in the hallways at professional meetings. Accompanying this Declaration as Exhibit NNN is a true copy of advertisements promoting the 1999 Standards, marked as Exhibit 1218 during my deposition.

15. All copies of the Standards bear a copyright notice.

16. Distribution of the Standards is closely monitored by the Sponsoring Organizations. AERA, the designated publisher of the Standards, sometimes provides promotional complementary print copies to students or professors. Except for these few complementary print copies, however, the Standards are not given away for free; and certainly

they are not made available to the public by any of the three organizations for anyone to copy free of charge. To date, AERA has never posted, or authorized the posting of, a digitized copy of the 1999 Standards on any publicly accessible website.

17. The 1999 Standards have been sold at retail prices ranging from \$25.95 to \$49.95 per copy. From 2000 to 2014, except for the near two-year period during which Public Resource posted unauthorized copies online and sales diminished significantly, income generated from sales of the 1999 Standards, on average, had been approximately in excess of \$127,000 per year.

18. Accompanying this Declaration as Exhibit OOO is a true copy of AERA's Statement of Revenue and Expenses for the Standards from FY2000 to December 31, 2013, marked as Exhibit 1211 during my deposition.

19. After the 2014 Standards were published in the late summer of 2014, AERA for a time discontinued sales of the 1999 Standards. This was to encourage sales of the newly-revised edition – the 2014 Standards. Accompanying this Declaration as Exhibit PPP is a true copy of the publication page for the 1999 Standards on the AERA website as of May 4, 2015 showing that the 1999 Standards were not available for sale at that time, marked as Exhibit 1196 during my deposition.

20. However, so long as purchasers are made aware that it is no longer the current edition, the 1999 Standards do have an enduring value for those in the testing and assessment profession who (i) need to know the state of best testing practices as they existed between 1999 and 2014, (ii) believe they still may be held accountable to the guidance of the 1999 Standards even now, and/or (iii) study the changes in best testing and assessment practices over time. For this reason, in the summer of 2015 AERA resumed sales of the 1999 Standards. Accompanying this Declaration as Exhibit QQQ is a true copy of the publication page for the 1999 Standards on

the AERA website as updated during the summer of 2015, showing that the 1999 Standards are available for sale.

21. All revenue from the sale of the 1999 Standards above expenses is used to cover the publishing costs of the Standards and for the preparation of subsequent editions of the Standards. The Sponsoring Organizations do not distribute any proceeds from the sales of the Standards to the Sponsoring Organizations. Rather, the income from these sales is used by the Sponsoring Organizations to offset their development and production costs and to generate funds for subsequent revisions. This allows the Sponsoring Organizations to develop up-to-date, high quality Standards that otherwise would not be developed due to the time and effort that goes into producing them.

22. Without receiving revenue from the sales of the Standards to offset their preparation costs and to allow for further revisions, it is very likely that the Sponsoring Organizations would no longer undertake to periodically update them, and it is unknown who else would.

23. The Sponsoring Organizations decided on a model of self-funding of revisions of the Standards; that is, from the sale of prior editions of the Standards. Funding for the Standards revision process from third party sources (*e.g.*, governmental agencies, foundations, other associations interested in testing and assessment issues, etc.) was rejected because of the appearance or potential of conflicts of interest and the importance of users of the Standards being able to trust in their scientific integrity.

24. Due to the relative minor portion of the membership of AERA who devote their careers to testing and assessment, it is highly unlikely that the members of AERA will vote for a dues increase to fund future Standards revision efforts if Public Resource successfully defends

this case and is allowed to post the Standards online for the public to download or print for free. As a result, the Sponsoring Organizations would likely abandon their practice of periodically updating the Standards.

25. The Standards were registered with the U.S. Register of Copyrights under Registration Number TX 5-100-196, having an effective date of December 8, 1999. Accompanying this Declaration as Exhibit RRR is a true copy of the December 8, 1999 Copyright Certificate of Registration for the 1999 Standards.

26. A supplementary copyright registration for the Standards was issued by the U.S. Register of Copyrights under Supplementary Registration Number TX 6-434-609, having an effective date of February 25, 2014. This Supplementary Registration was obtained to correct an error in the listing of copyright ownership in Registration Number TX 5-100-196. Accompanying this Declaration as Exhibit SSS is a true copy of the February 25, 2014 Supplementary Copyright Certificate of Registration for the 1999 Standards.

27. The Joint Committee that authored the 1999 Standards comprised 16 members.

28. Accompanying this Declaration as Exhibit TTT is a true copy of the 1999 Standards.

29. Public Resource posted Plaintiffs' 1999 Standards to its website and the Internet Archive website without the permission or authorization of any of the Sponsoring Organizations.

30. The Sponsoring Organizations can only speculate on the number of electronic copies of the 1999 Standards that were made and distributed to others by the original Internet users who accessed the unauthorized copies that Public Resource posted to its site and the Internet Archive site. There simply is no way for the Sponsoring Organizations to calculate with any degree of certainty the number of university/college professors, students, testing companies

and others who would have purchased Plaintiffs' Standards but for their wholesale posting on Defendant's <https://law.resource.org> website and the Internet Archive <http://archive.org> website.

31. In December 2013, Plaintiff AERA requested in writing that Public Resource remove the 1999 Standards from its online postings. Accompanying this Declaration as Exhibit UUU is a true copy of a letter sent from John S. Neikirk, Director of Publications at AERA, to Carl Malamud of Public Resource regarding the posting of the 1999 Standards at <https://law.resource.org/pub/us/cfr/ibr/001/aera.standards.1999.pdf>, marked as Exhibit 1228 during my deposition.

32. Had Public Resource not promised to remove the 1999 Standards from its law.resource.org website and the Internet Archive website while this lawsuit is pending, and followed through with, these promises, the Sponsoring Organizations seriously contemplated moving forward with a motion to preliminary enjoin Public Resource from maintaining the unauthorized postings of electronic copies of the 1999 Standards on the Internet, and delaying publication of the 2014 Standards.

33. By June 2014, when Public Resource finally removed its online postings of the 1999 Standards, the damage already had been done. In Fiscal Year ("FY") 2011 to FY 2012, as compared to FY 2011, the Sponsoring Organizations experienced a 34% drop in sales of the 1999 Standards. In FY 2013, sales of the 1999 Standards remained at their low level from the prior fiscal year.

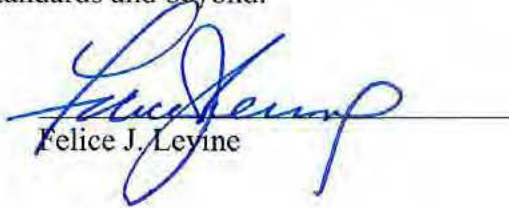
34. This is notable, given that Public Resource posted the Standards to the Internet in 2012-2013, and that the Sponsoring Organizations' updated Standards were not published until the summer of 2014.

35. Past harm from Public Resource's infringing activities includes lost sales that

cannot be totally accounted for – due to potentially infinite Internet distribution; for example, by psychometrics students – and a lack of funding that otherwise would have been available for the update of the Sponsoring Organizations’ Standards from the 1999 to the 2014 versions.

36. Should Public Resource’s infringement be allowed to continue, the harm to the Sponsoring Organizations, and public at large who rely on the preparation and administration of valid, fair and reliable tests, includes: (i) uncontrolled publication of the 1999 Standards without any notice that those guidelines have been replaced by the 2014 Standards; (ii) future unquantifiable loss of revenue from sales of authorized copies of the 1999 Standards (with proper notice that they are no longer the current version) and the 2014 Standards; and (iii) lack of funding for future revisions of the 2014 Standards and beyond.

Dated: December 15, 2015



Felice J. Levine

EXHIBIT QQQ

Case No. 1:14-cv-00857-TSC-DAR



[Graduate Students](#) | [Divisions](#) | [SIGs](#) | [OIA](#)



[Publications](#) » [Books](#) » [Standards for Educational & Psychological Testing](#)

Standards for Educational & Psychological Testing

Journals

AERA Highlights

Books

Research Points

Online Paper Repository

Online Store

Advertise with AERA

Publications Permissions



Revised significantly from the 1985 version, the 1999 Standards has more in-depth background material in each chapter, a greater number of standards, and a significantly expanded glossary and index. The new Standards reflects changes in federal law and measurement trends affecting validity; testing individuals with disabilities or different linguistic backgrounds; and new types of tests as well as new uses of existing tests. The Standards is written for the professional and for the educated layperson and addresses professional and technical issues of test development and use in education, psychology, and employment. This book is a vitally important reference for professional test developers, sponsors, publishers, users, policymakers, employers, and students in education and psychology.

Revised 1999. Paperback ([order form](#))
Developed jointly by the American Educational Research Association, American Psychological Association, and the National Council on [Measurement in Education](#)

AERA released a [new edition](#) of the Testing Standards in 2014.

©2015 American Educational Research Association. All rights reserved.

[Terms Of Use](#) | [Privacy Policy](#) | [Site Map](#) | [Contact Us](#)

1430 K Street NW, Suite 1200, Washington, DC 20005
Phone: (202) 238-3200 | Fax: (202) 238-3250

Designed by [Weber-Shandwick](#) Powered by [eNOAH](#)

EXHIBIT RRR

Case No. 1:14-cv-00857-TSC-DAR

FORM TX
 UNITED STATES COPYRIGHT OFFICE

REGISTRATION NUMBER

107028275

TX 5-100-196

(TX) TXU
 EFFECTIVE DATE OF REGISTRATION
 12 8 99
 Month Day Year

DO NOT WRITE ABOVE THIS LINE. IF YOU NEED MORE SPACE, USE A SEPARATE CONTINUATION SHEET.

1

TITLE OF THIS WORK ▼

Standards for Educational and Psychological Testing

PREVIOUS OR ALTERNATIVE TITLES ▼

N/A

PUBLICATION AS A CONTRIBUTION If this work was published as a contribution to a periodical, serial, or collection, give information about the collective work in which the contribution appeared. **Title of Collective Work** ▼

N/A

If published in a periodical or serial give: **Volume** ▼ **Number** ▼ **Issue Date** ▼ **On Pages** ▼
 N/A

2

NAME OF AUTHOR ▼

American Educational Research Association

DATES OF BIRTH AND DEATH
 Year Born ▼ Year Died ▼
 N/A N/A

Was this contribution to the work a "work made for hire"?
 Yes No
AUTHOR'S NATIONALITY OR DOMICILE
 Name of Country: USA
 OR Citizen of: _____
 Domiciled in: _____
WAS THIS AUTHOR'S CONTRIBUTION TO THE WORK
 Anonymously? Yes No
 Pseudonymously? Yes No
 If the answer to either of these questions is "Yes," see detailed instructions.

NATURE OF AUTHORSHIP Briefly describe nature of the material created by this author in which copyright is claimed. ▼

NOTE

Under the law, the "author" of a "work made for hire" is generally the employer, not the employee (see instructions). For any part of this work that was "made for hire," check "Yes" in the space provided, give the employer (or other person for whom the work was prepared) as "Author" of that part, and leave the space for dates of birth and death blank.

NAME OF AUTHOR ▼

b

Was this contribution to the work a "work made for hire"?
 Yes No
AUTHOR'S NATIONALITY OR DOMICILE
 Name of Country: _____
 OR Citizen of: _____
 Domiciled in: _____
WAS THIS AUTHOR'S CONTRIBUTION TO THE WORK
 Anonymously? Yes No
 Pseudonymously? Yes No
 If the answer to either of these questions is "Yes," see detailed instructions.

NATURE OF AUTHORSHIP Briefly describe nature of the material created by this author in which copyright is claimed. ▼

NAME OF AUTHOR ▼

c

Was this contribution to the work a "work made for hire"?
 Yes No
AUTHOR'S NATIONALITY OR DOMICILE
 Name of Country: _____
 OR Citizen of: _____
 Domiciled in: _____
WAS THIS AUTHOR'S CONTRIBUTION TO THE WORK
 Anonymously? Yes No
 Pseudonymously? Yes No
 If the answer to either of these questions is "Yes," see detailed instructions.

NATURE OF AUTHORSHIP Briefly describe nature of the material created by this author in which copyright is claimed. ▼

3

YEAR IN WHICH CREATION OF THIS WORK WAS COMPLETED This information must be given in all cases. 1999
DATE AND NATION OF FIRST PUBLICATION OF THIS PARTICULAR WORK Complete this information ONLY if this work has been published. Month: November Day: 8 Year: 1999
 USA

4

COPYRIGHT CLAIMANT(S) Name and address must be given even if the claimant is the same as the author given in space 2. ▼

American Educational Research Association
 1230 17th St., NW
 Washington, DC 20036

APPLICATION RECEIVED
 DEC 08 1999
 ONE DEPOSIT RECEIVED
 TWO DEPOSITS RECEIVED
 DEC 08 1999
 REMITTANCE NUMBER AND DATE

See instructions before completing this space.

TRANSFER If the claimant(s) named here in space 4 are different from the author(s) named in space 2, give a brief statement of how the claimant(s) obtained ownership of the copyright. ▼

MORE ON BACK ▶ Complete all applicable spaces (numbers 5-11) on the reverse side of this page. See detailed instructions. Sign the form at line 10.

DO NOT WRITE HERE
 Page 1 of 2

JA2591

EXAMINED BY NO FORM TX
CHECKED BY _____
 CORRESPONDENCE Yes
 DEPOSIT ACCOUNT FUNDS USED
FOR COPYRIGHT OFFICE USE ONLY

DO NOT WRITE ABOVE THIS LINE. IF YOU NEED MORE SPACE, USE A SEPARATE CONTINUATION SHEET.

PREVIOUS REGISTRATION Has registration for this work, or for an earlier version of this work, already been made in the Copyright Office?

- Yes No If your answer is "Yes," why is another registration being sought? (Check appropriate box) ▼
- This is the first published edition of a work previously registered in unpublished form.
- This is the first application submitted by this author as copyright claimant.
- This is a changed version of the work, as shown by space 6 on this application.

If your answer is "Yes," give: Previous Registration Number ▼ Year of Registration ▼

5

DERIVATIVE WORK OR COMPILATION Complete both space 6a & 6b for a derivative work; complete only 6b for a compilation.

a. Preexisting Material Identify any preexisting work or works that this work is based on or incorporates. ▼

6

b. Material Added to This Work Give a brief, general statement of the material that has been added to this work and in which copyright is claimed. ▼

See instructions before completing this space.

MANUFACTURERS AND LOCATIONS If this is a published work consisting preponderantly of nondramatic literary material in English, the law may require that the copies be manufactured in the United States or Canada for full protection. If so, the names of the manufacturers who performed certain processes, and the places where these processes were performed must be given. See instructions for details.

Names of Manufacturers ▼ Places of Manufacture ▼

7

REPRODUCTION FOR USE OF BLIND OR PHYSICALLY HANDICAPPED INDIVIDUALS A signature on this form at space 10, and a check in one of the boxes here in space 8, constitutes a non-exclusive grant of permission to the Library of Congress to reproduce and distribute solely for the blind and physically handicapped and under the conditions and limitations prescribed by the regulations of the Copyright Office: (1) copies of the work identified in space 1 of this application in Braille (or similar tactile symbols); or (2) phonorecords embodying a fixation of a reading of that work; or (3) both.

- a Copies and Phonorecords
- b Copies Only
- c Phonorecords Only

See instructions.

DEPOSIT ACCOUNT If the registration fee is to be charged to a Deposit Account established in the Copyright Office, give name and number of Account.

Name ▼ Account Number ▼
American Educational Research Association DAO 31143

8

9

CORRESPONDENCE Give name and address to which correspondence about this application should be sent. Name/Address/Apt/City/State/Zip ▼

Camille S. Coy
1230 17th St., NW
Washington, DC 20036

Area Code & Telephone Number ▶ 202-223-9485

Be sure to give your daytime phone number

CERTIFICATION* I, the undersigned, hereby certify that I am the

Check one ▶

- author
- other copyright claimant
- owner of exclusive right(s)
- authorized agent of AERA

of the work identified in this application and that the statements made by me in this application are correct to the best of my knowledge. Name of author or other copyright claimant, or owner of exclusive right(s) ▶

Typed or printed name and date ▼ If this is a published work, this date must be the same as or later than the date of publication given in space 3.

Camille S. Coy date ▶ 12/3/99

10

Handwritten signature (X) ▼

MAIL CERTIFICATE TO

Certificate will be mailed in window envelope

Name ▼ American Educational Research Association
Number/Street/Apartment Number ▼ 1230 17th St., NW
City/State/ZIP ▼ Washington, DC 20036

- Have you:
- Completed all necessary spaces?
 - Signed your application in space 10?
 - Enclosed check or money order for \$10 payable to Register of Copyrights?
 - Enclosed your deposit material with the application and fee?
- MAIL TO: Register of Copyrights, Library of Congress, Washington, D.C. 20559.

11

* 17 U.S.C. § 506(e): Any person who knowingly makes a false representation of a material fact in the application for copyright registration provided for by section 409, or in any written statement filed in connection with the application, shall be fined not more than \$2,500.

EXHIBIT SSS

Case No. 1:14-cv-00857-TSC-DAR

Certificate of Registration



This Certificate issued under the seal of the Copyright Office in accordance with title 17, *United States Code*, attests that registration has been made for the work identified below. The information on this certificate has been made a part of the Copyright Office records.

Maria A. Pallante

Register of Copyrights, United States of America

Form CA
For Supplementary Registration
UNITED STATES COPYRIGHT OFFICE

TX 6-484-609



TX TXU PA PAU VA VAU BR BRU RS
EFFECTIVE DATE OF SUPPLEMENTARY REGISTRATION

2 25 2014
Month Day Year

DO NOT WRITE ABOVE THIS LINE. IF YOU NEED MORE SPACE, USE A SEPARATE CONTINUATION SHEET.

A

Title of Work Standards for Educational and Psychological Testing

Registration Number of the Basic Registration
TX 5-100-198

Year of Basic Registration
1999

Name(s) of Author(s)
American Educational Research Association

Name(s) of Copyright Claimant(s)
American Educational Research Association

B

Location and Nature of Incorrect Information in Basic Registration
Line Number 2b Line Heading or Description Name of Author

Incorrect Information as it Appears in Basic Registration
Blank

Corrected Information
American Psychological Association

Explanation of Correction
This book had three authors, but only one was included in the original filing.

C

Location and Nature of Information in Basic Registration to be Amplified
Line Number 3a Line Heading or Description Copyright Claimant(s)

Amplified Information and Explanation of Information
The address of the claimant listed has changed to:

American Educational Research Association
1430 K Street NW, Suite 1200
Washington, DC 20005

MORE ON BACK • Complete all applicable spaces (D-G) on the reverse side of this page.
• See detailed instructions. • Sign the form at Space F.

DO NOT WRITE HERE
Page 1 of 2 pages

FORM CA RECEIVED

FORM CA

FUNDS RECEIVED DATE

EXAMINED BY

KOKA

CORRESPONDENCE

REFERENCE TO THIS REGISTRATION ADDED TO BASIC REGISTRATION YES NO

FOR COPYRIGHT OFFICE USE ONLY

DO NOT WRITE ABOVE THIS LINE. IF YOU NEED MORE SPACE, USE A SEPARATE CONTINUATION SHEET.

Continuation of Part B or Part C

Line 2c was blank, but should have included the third author: National Council on Measurement in Education

Line 4 listed one copyright claimant: American Educational Research Association. Line 4 should have listed 3 copyright claimants as listed below:

American Educational Research Association
1430 K Street, NW
Suite 1200
Washington, DC 20005

American Psychological Association
750 First Street NE
Washington, DC 20002-4242

National Council on Measurement in Education
2424 American Lane
Madison, WI 53704

D

Correspondence: Give name and address to which correspondence about this application should be sent.

John Nelkirk, 1430 K Street, NW
Suite 1200
Washington, DC 20005

Phone (202) 238-3238

Fax (202) 238-3250

Email jnelkirk@aera.net

Deposit Account: If the registration fee is to be charged to a Deposit Account established in the Copyright Office, give name and number of Account.

Name _____
Account Number _____

Certification* I, the undersigned, hereby certify that I am the: (Check only one)

- author
- owner of exclusive right(s)
- other copyright claimant
- duly authorized agent of

American Educational Research Association

Name of author or other copyright claimant, or owner of exclusive right(s) & of the work identified in this application and that the statements made by me in this application are correct to the best of my knowledge.

Typed or printed name John Nelkirk

Date 2/24/2014

Handwritten signature

John Nelkirk

Certificate will be mailed in window envelope to this address:

Name John Nelkirk, American Educational Research Association

Number/Street/Apt 1430 K Street NW, Suite 1200

City/State/Zip Washington, DC 20005

RECEIVED

U.S. Copyright Office

101 Independence Avenue SE
Washington, DC 20540-4400

*17 USC 405: Any person who knowingly makes a false representation of a material fact in the application for copyright registration provided for by section 405, or in any written statement filed in connection with the application, shall be fined not more than \$5,000.

EXHIBIT TTT-1

Case No. 1:14-cv-00857-TSC-DAR

STANDARDS

for the Practice of Psychology

American Psychological Association
American Psychological Association
National Council on Accreditation of Counseling

Standards for Educational and Psychological Testing **AERA, APA, NCME**

JA2598

STANDARDS

for educational and psychological testing

American Educational Research Association
American Psychological Association
National Council on Measurement in Education

Copyright © 1999 by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. All rights reserved. Except as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

Published by
American Educational Research Association
1430 K St., NW, Suite 1200
Washington, DC 20005

Library of Congress Card number: 99066845
ISBN: 0-935302-25-5
ISBN-13: 978-0-935302-25-7

Printed in the United States of America
First printing in 1999; second, 2002; third, 2004;
fourth, 2007; fifth, 2008; and sixth, 2011.

The *Standards for Educational and Psychological Testing* will be under continuing review by the three sponsoring organizations. Comments and suggestions will be welcome and should be sent to The Committee to Develop Standards for Educational and Psychological Testing in care of the Executive Office, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

Prepared by the
Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.

TABLE OF CONTENTS

PREFACE	v
INTRODUCTION	1
Participants in the Testing Process	1
The Purpose of the Standards	2
Categories of Standards	2
Tests and Test Uses to Which These Standards Apply	3
Cautions to be Exercised in Using the Standards	4
The Number of Standards	4
Tests as Measures of Constructs	5
Organization of This Volume	5
PART I	
TEST CONSTRUCTION, EVALUATION, AND DOCUMENTATION	7
1. Validity	9
Background	9
Standards 1.1-1.24	17
2. Reliability and Errors of Measurement	25
Background	25
Standards 2.1-2.20	31
3. Test Development and Revision	37
Background	37
Standards 3.1-3.27	43
4. Scales, Norms, and Score Comparability	49
Background	49
Standards 4.1-4.21	54
5. Test Administration, Scoring, and Reporting	61
Background	61
Standards 5.1-5.16	63
6. Supporting Documentation for Tests	67
Background	67
Standards 6.1-6.15	68
PART II	
FAIRNESS IN TESTING	71
7. Fairness in Testing and Test Use	73
Background	73
Standards 7.1-7.12	80

TABLE OF CONTENTS

8. The Rights and Responsibilities of Test Takers	85
Background	85
Standards 8.1-8.13	86
9. Testing Individuals of Diverse Linguistic Backgrounds	91
Background	91
Standards 9.1-9.11	97
10. Testing Individuals with Disabilities	101
Background	101
Standards 10.1-10.12	106
PART III	
TESTING APPLICATIONS	109
11. The Responsibilities of Test Users	111
Background	111
Standards 11.1-11.24	113
12. Psychological Testing and Assessment	119
Background	119
Standards 12.1-12.20	131
13. Educational Testing and Assessment	137
Background	137
Standards 13.1-13.19	145
14. Testing in Employment and Credentialing	151
Background	151
Standards 14.1-14.17	158
15. Testing in Program Evaluation and Public Policy	163
Background	163
Standards 15.1-15.13	167
GLOSSARY	171
INDEX	185

PREFACE

There have been five earlier documents from three sponsoring organizations guiding the development and use of tests. The first of these was *Technical Recommendations for Psychological Tests and Diagnostic Techniques*, prepared by a committee of the American Psychological Association (APA) and published by that organization in 1954. The second was *Technical Recommendations for Achievement Tests*, prepared by a committee representing the American Educational Research Association (AERA) and the National Council on Measurement Used in Education (NCMUE) and published by the National Education Association in 1955. The third, which replaced the earlier two, was published by APA in 1966 and prepared by a committee representing APA, AERA, and the National Council on Measurement in Education (NCME) and called the *Standards for Educational and Psychological Tests and Manuals*. The fourth, *Standards for Educational and Psychological Tests*, was again a collaboration of AERA, APA and NCME, and was published in 1974. The fifth, *Standards for Educational and Psychological Testing*, also a joint collaboration, was published in 1985.

In 1991 APA's Committee on Psychological Tests and Assessment suggested the need to revise the 1985 *Standards*. Representatives of AERA, APA and NCME met and discussed the revision, principles that should guide that revision, and potential Joint Committee members. By 1993, the presidents of the three organizations appointed members and the Committee had its first meeting November, 1993.

The *Standards* has been developed by a joint committee appointed by AERA, APA and NCME. Members of the Committee were:

Eva Baker, *co-chair*
Paul Sackett, *co-chair*
Lloyd Bond
Leonard Feldt

David Goh
Bert Green
Edward Haertel
Jo-Ida Hansen
Sharon Johnson-Lewis
Suzanne Lane
Joseph Matarazzo
Manfred Meier
Pamela Moss
Esteban Olmedo
Diana Pullin

From 1993 to 1996 Charles Spielberger served on the Committee as co-chair. Each sponsoring organization was permitted to assign up to two liaisons to the Joint Committee's project. Liaisons served as the conduits between the sponsoring organizations and the Joint Committee. APA's liaison from its Committee on Psychological Tests and Assessments changed several times as the membership of the Committee changed.

Liaisons to the Joint Committee:

AERA - William Mehrens
APA - Bruce Bracken, Andrew Czopek,
Rodney Lowman, Thomas Oakland
NCME - Daniel Eignor

APA and NCME also had committees who served to monitor the process and keep relevant parties informed.

APA Ad Hoc Committee of the Council of Representatives:

Melba Vasquez
Donald Bersoff
Stephen DeMers
James Farr
Bertram Karon
Nadine Lambert
Charles Spielberger

NCME Standards and Test Use Committee:

Gregory Cizek
Allen Doolittle
Le Ann Gamache

Donald Ross Green
Ellen Julian
Tracy Muenz
Nambury Raju

A management committee was formed at the beginning of this effort. They monitored the financial and administrative arrangements of the project, and advised the sponsoring organizations on such matters.

Management Committee:

Frank Farley, APA
George Madaus, AERA
Wendy Yen, NCME

Staffing for the revision included Dianne Brown Maranto as project director, and Dianne L. Schneider as staff liaison. Wayne J. Camara served as project director from 1993 to 1994. APA's legal counsel conducted the legal review of the *Standards*. William C. Howell and William Mehrens reviewed the standards for consistency across chapters. Linda Murphy developed the indexing for the book.

The Joint Committee solicited preliminary reviews of some draft chapters, from recognized experts. These reviews were primarily solicited for the technical and fairness chapters. Reviewers are listed below:

Marvin Alkin
Philip Bashook
Bruce Bloxom
Jeffery P. Braden
Robert L. Brennan
John Callender
Ronald Cannella
Lee J. Cronbach
James Cummins
John Fremer
Kurt F. Geisinger
Robert M. Guion
Walter Haney
Patti L. Harrison
Gerald P. Koocher
Richard Jeanneret

Frank Landy
Ellen Lent
Robert Linn
Theresa C. Liu
Stanford von Mayrhauser
Milbrey W. McLaughlin
Samuel Messick
Craig N. Mills
Robert J. Mislavy
Kevin R. Murphy
Mary Anne Nester
Maria Pennock-Roman
Carole Perlman
Michael Rosenfeld
Jonathan Sandoval
Cynthia B. Schmeiser
Kara Schmitt
Neal Schmitt
Richard J. Shavelson
Lorrie A. Shepard
Mark E. Swerdlik
Janet Wall
Anthony R. Zara

Draft versions of the *Standards* were widely distributed for public review and comment three times during this revision effort, providing the Committee with a total of nearly 8,000 pages of comments. Organizations who submitted comments on drafts are listed below. Many individuals contributed to the input from each organization, and although we wish we could acknowledge every individual who had input, we cannot do so due to incomplete information as to who contributed to each organization's response. The Joint Committee could not have completed its task without the thoughtful reviews of so many professionals.

Sponsoring Associations

American Educational Research
Association (AERA)
American Psychological Association (APA)
National Council on Measurement in
Education (NCME)

PREFACE

Membership Organizations (Scientific, Professional, Trade & Advocacy)

American Association for Higher Education (AAHE)
American Board of Medical Specialties (ABMS)
American Counseling Association (ACA)
American Evaluation Association (AEA)
American Occupational Therapy Association
American Psychological Society (APS)
APA Division of Counseling Psychology (Division 17)
APA Division of Developmental Psychology (Division 7)
APA Division of Evaluation, Measurement, and Statistics (Division 5)
APA Division of Mental Retardation & Developmental Disabilities (Division 33)
APA Division of Pharmacology & Substance Abuse (Division 28)
APA Division of Rehabilitation Psychology (Division 22)
APA Division of School Psychology (Division 16)
Asian American Psychological Association (AAPA)
Association for Assessment in Counseling (AAC)
Association of Test Publishers (ATP)
Australian Council for Educational Research Limited (ACER)
Chicago Industrial/Organizational Psychologists (CIOP)
Council on Licensure, Enforcement, and Regulation (CLEAR), Examination Resources & Advisory Committee (ERAC)
Equal Employment Advisory Council (EEAC)
Foundation for Rehabilitation Certification, Education and Research
Human Sciences Research Council, South Africa
International Association for Cross-Cultural Psychology (IACCP)

International Brotherhood of Electrical Workers
International Language Testing Association
International Personnel Management Association Assessment Council (IPMAAC)
Joint Committee on Testing Practices (JCTP)
National Association for the Advancement of Colored People (NAACP), Legal Defense and Educational Fund, Inc.
National Center for Fair and Open Testing (Fairtest)
National Organization for Competency Assurance (NOCA)
Personnel Testing Council of Metropolitan Washington (PTC/MW)
Personnel Testing Council of Southern California (PTC/SC)
Society for Human Resource Management (SHRM)
Society of Indian Psychologists (SIP)
Society for Industrial and Organizational Psychology (APA Division 14)
Society for the Psychological Study of Ethnic Minority Issues (APA Division 45)
State Collaborative on Assessment & Student Standards Technical Guidelines for Performance Assessment Consortium (TGPA)
Telecommunications Staffing Forum
Western Region Intergovernmental Personnel Assessment Council (WRIPAC)

Credentialing Boards

American Board of Physical and Medical Rehabilitation
American Medical Technologists
Commission on Rehabilitation Counselor Certification
National Board for Certified Counselors (NBCC)
National Board of Examiners in Optometry

National Board of Medical Examiners
National Council of State Boards of
Nursing

Government and Federal Agencies

Army Research Institute (ARI)
California Highway Patrol, Personnel and
Training Division, Selection Research
Program
City of Dallas, Civil Service Department
Commonwealth of Virginia, Department
of Education
Defense Manpower Data Center
(DMDC), Personnel Testing Division
Department of Defense (DOD), Office
of the Assistant Secretary of Defense
Department of Education, Office of
Educational Improvement, National
Center for Education Statistics
Department of Justice, Immigration and
Naturalization Service (INS)
Department of Labor, Employment and
Training Administration (DOL/ETA)
U.S. Equal Employment Opportunity
Commission (EEOC)
U.S. Office of Personnel Management
(OPM), Personnel Resources &
Development Center

Test Publishers/Developers

American College Testing (ACT)
CTB/McGraw-Hill
The College Board
Educational Testing Service (ETS)
Highland Publishing Company
Institute for Personality & Ability
Testing (IPAT)
Professional Examination Service (PES)

Academic Institutions

Center for Creative Leadership
Gallaudet University, National Task
Force on Equity in Testing Deaf
Professionals
University of Haifa, Israeli Group
Kansas State University
National Center on Educational
Outcomes (NCEO)

Pennsylvania State University
University of North Carolina – Charlotte
University of Southern Mississippi,
Department of Psychology

When the Joint Committee completed its task of revising the *Standards*, it then submitted its work to the three sponsoring organizations for approval. Each organization had its own governing body and mechanism for approval, as well as definitions for what their approval means.

AERA: This endorsement carries with it the understanding that, in general, we believe the *Standards* to represent the current consensus among recognized professionals regarding expected measurement practice. Developers, sponsors, publishers, and users of tests should observe these *Standards*.

APA: The APA's approval of the *Standards* means the Council adopts the document as APA policy.

NCME: NCME endorses the *Standards for Educational and Psychological Testing* and recognizes that the intent of these *Standards* is to promote sound and responsible measurement practice. This endorsement carries with it a professional imperative for NCME members to attend to the *Standards*.

Although the *Standards* are prescriptive, the *Standards* itself does not contain enforcement mechanisms. These standards were formulated with the intent of being consistent with other standards, guidelines and codes of conduct published by the three sponsoring organizations, and listed below. The reader is encouraged to obtain these documents, some of which have references to testing and assessment in specific applications or settings.

The Joint Committee on the
*Standards for Educational and
Psychological Testing*

PREFACE

References

American Educational Research Association. (June, 1992). *Ethical Standards of the American Educational Research Association*. Washington, DC: Author.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. *Standards for Teacher Competence in Educational Assessment of Students*. (1990). Washington, DC: National Council on Measurement in Education.

American Psychological Association. (December, 1992). Ethical Principles of Psychologists and Code of Conduct. *American Psychologist*, 47 (12), 1597-1611.

Joint Committee on Testing Practices. (1988). *Code of Fair Testing Practices in Education*. Washington, DC: American Psychological Association.

National Council on Measurement in Education. (1995). *Code of Professional Responsibilities in Educational Measurement*. Washington, DC: Author.

INTRODUCTION

Educational and psychological testing and assessment are among the most important contributions of behavioral science to our society, providing fundamental and significant improvements over previous practices. Although not all tests are well-developed nor are all testing practices wise and beneficial, there is extensive evidence documenting the effectiveness of well-constructed tests for uses supported by validity evidence. The proper use of tests can result in wiser decisions about individuals and programs than would be the case without their use and also can provide a route to broader and more equitable access to education and employment. The improper use of tests, however, can cause considerable harm to test takers and other parties affected by test-based decisions. The intent of the *Standards* is to promote the sound and ethical use of tests and to provide a basis for evaluating the quality of testing practices.

Participants in the Testing Process

Educational and psychological testing and assessment involve and significantly affect individuals, institutions, and society as a whole. The individuals affected include students, parents, teachers, educational administrators, job applicants, employees, clients, patients, supervisors, executives, and evaluators, among others. The institutions affected include schools, colleges, businesses, industry, clinics, and government agencies. Individuals and institutions benefit when testing helps them achieve their goals. Society, in turn, benefits when testing contributes to the achievement of individual and institutional goals.

The interests of the various parties involved in the testing process are usually, but not always, congruent. For example, when a test is given for counseling purposes or for job placement, the interests of the individual and the institution often coincide. In contrast, when a test is used to

select from among many individuals for a highly competitive job or for entry into an educational or training program, the preferences of an applicant may be inconsistent with those of an employer or admissions officer. Similarly, when testing is mandated by a court, the interests of the test taker may be different from those of the party requesting the court order.

There are many participants in the testing process, including, among others: (a) those who prepare and develop the test; (b) those who publish and market the test; (c) those who administer and score the test; (d) those who use the test results for some decision-making purpose; (e) those who interpret test results for clients; (f) those who take the test by choice, direction, or necessity; (g) those who sponsor tests, which may be boards that represent institutions or governmental agencies that contract with a test developer for a specific instrument or service; and (h) those who select or review tests, evaluating their comparative merits or suitability for the uses proposed.

These roles are sometimes combined and sometimes further divided. For example, in clinics the test taker is typically the intended beneficiary of the test results. In some situations the test administrator is an agent of the test developer, and sometimes the test administrator is also the test user. When an industrial organization prepares its own employment tests, it is both the developer and the user. Sometimes a test is developed by a test author but published, advertised, and distributed by an independent publisher, though the publisher may play an active role in the test development. Given this intermingling of roles, it is difficult to assign precise responsibility for addressing various standards to specific participants in the testing process.

This document begins with a series of chapters on the test development process, which focus primarily on the responsibilities of test developers, and then turns to chapters

on specific uses and applications, which focus primarily on responsibilities of test users. One chapter is devoted specifically to the rights and responsibilities of test takers.

The *Standards* is based on the premise that effective testing and assessment require that all participants in the testing process possess the knowledge, skills, and abilities relevant to their role in the testing process, as well as awareness of personal and contextual factors that may influence the testing process. They also should obtain any appropriate supervised experience and legislatively mandated practice credentials necessary to perform competently those aspects of the testing process in which they engage. For example, test developers and those selecting and interpreting tests need adequate knowledge of psychometric principles such as validity and reliability.

The Purpose of the Standards

The purpose of publishing the *Standards* is to provide criteria for the evaluation of tests, testing practices, and the effects of test use. Although the evaluation of the appropriateness of a test or testing application should depend heavily on professional judgment, the *Standards* provides a frame of reference to assure that relevant issues are addressed. It is hoped that all professional test developers, sponsors, publishers, and users will adopt the *Standards* and encourage others to do so.

The *Standards* makes no attempt to provide psychometric answers to questions of public policy regarding the use of tests. In general, the *Standards* advocates that, within feasible limits, the relevant technical information be made available so that those involved in policy debate may be fully informed.

Categories of Standards

The 1985 *Standards* designated each standard as "primary" (to be met by all tests before operational use), "secondary" (desirable, but

not feasible in certain situations), or "conditional" (importance varies with application). The present *Standards* continues the tradition of expecting test developers and users to consider all standards before operational use; however, the *Standards* does not continue the practice of designating levels of importance. Instead, the text of each standard, and any accompanying commentary, discusses the conditions under which a standard is relevant. It was not the case that under the 1985 *Standards* test developers and users were obligated to attend only to the primary standards. Rather, the term "conditional" meant that a standard was primary in some settings and secondary in others, thus requiring careful consideration of the applicability of each standard for a given setting.

The absence of designations such as "primary" or "conditional" should not be taken to imply that all standards are equally significant in any given situation. Depending on the context and purpose of test development or use, some standards will be more salient than others. Moreover, some standards are broad in scope, setting forth concerns or requirements relevant to nearly all tests or testing contexts, and other standards are narrower in scope. However, all standards are important in the contexts to which they apply. Any classification that gives the appearance of elevating the general importance of some standards over others could invite neglect of some standards that need to be addressed in particular situations.

Further, the current *Standards* does not include standards considered secondary or "desirable." The continued use of the secondary designation would risk encouraging both the expansion of the *Standards* to encompass large numbers of "desirable" standards and the inappropriate assumption that any guideline not included in the *Standards* as at least "secondary" was inconsequential.

Unless otherwise specified in the standard or commentary, and with the caveats

INTRODUCTION

outlined below, standards should be met before operational test use. This means that each standard should be carefully considered to determine its applicability to the testing context under consideration. In a given case there may be a sound professional reason why adherence to the standard is unnecessary. It is also possible that there may be occasions when technical feasibility may influence whether a standard can be met prior to operational test use. For example, some standards may call for analyses of data that may not be available at the point of initial operational test use. If test developers, users, and, when applicable, sponsors have deemed a standard to be inapplicable or unfeasible, they should be able, if called upon, to explain the basis for their decision. However, there is no expectation that documentation be routinely available of the decisions related to each standard.

Tests and Test Uses to Which These Standards Apply

A test is an evaluative device or procedure in which a sample of an examinee's behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process. While the label *test* is ordinarily reserved for instruments on which responses are evaluated for their correctness or quality and the terms *scale* or *inventory* are used for measures of attitudes, interest, and dispositions, the *Standards* uses the single term *test* to refer to all such evaluative devices.

A distinction is sometimes made between *test* and *assessment*. *Assessment* is a broader term, commonly referring to a process that integrates test information with information from other sources (e.g., information from the individual's social, educational, employment, or psychological history). The applicability of the *Standards* to an evaluation device or method is not altered by the label applied to it (e.g., test, assessment, scale, inventory).

Tests differ on a number of dimensions: the mode in which test materials are presented (paper and pencil, oral, computerized administration, and so on); the degree to which stimulus materials are standardized; the type of response format (selection of a response from a set of alternatives as opposed to the production of a response); and the degree to which test materials are designed to reflect or simulate a particular context. In all cases, however, tests standardize the process by which test-taker responses to test materials are evaluated and scored. As noted in prior versions of the *Standards*, the same general types of information are needed for all varieties of tests.

The precise demarcation between those measurement devices used in the fields of educational and psychological testing that do and do not fall within the purview of the *Standards* is difficult to identify. Although the *Standards* applies most directly to standardized measures generally recognized as "tests," such as measures of ability, aptitude, achievement, attitudes, interests, personality, cognitive functioning, and mental health, it may also be usefully applied in varying degrees to a broad range of less formal assessment techniques. Admittedly, it will generally not be possible to apply the *Standards* rigorously to unstandardized questionnaires or to the broad range of unstructured behavior samples used in some forms of clinic- and school-based psychological assessment (e.g., an intake interview), and to instructor-made tests that are used to evaluate student performance in education and training. It is useful to distinguish between devices that lay claim to the concepts and techniques of the field of educational and psychological testing from those which represent nonstandardized or less standardized aids to day-to-day evaluative decisions. Although the principles and concepts underlying the *Standards* can be fruitfully applied to day-to-day decisions, such as when a business owner interviews a job applicant, a manager evalu-

ates the performance of subordinates, or a coach evaluates a prospective athlete, it would be overreaching to expect that the standards of the educational and psychological testing field be followed by those making such decisions. In contrast, a structured interviewing system developed by a psychologist and accompanied by claims that the system has been found to be predictive of job performance in a variety of other settings falls within the purview of the *Standards*.

Cautions to be Exercised in Using the Standards

Several cautions are important to avoid misinterpreting the *Standards*:

1) Evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every standard in this document, and acceptability cannot be determined by using a checklist. Specific circumstances affect the importance of individual standards, and individual standards should not be considered in isolation. Therefore, evaluating acceptability involves (a) professional judgment that is based on a knowledge of behavioral science, psychometrics, and the community standards in the professional field to which the tests apply; (b) the degree to which the intent of the standard has been satisfied by the test developer and user; (c) the alternatives that are readily available; and (d) research and experiential evidence regarding feasibility of meeting the standard.

2) When tests are at issue in legal proceedings and other venues requiring expert witness testimony it is essential that professional judgment be based on the accepted corpus of knowledge in determining the relevance of particular standards in a given situation. The intent of the *Standards* is to offer guidance for such judgments.

3) Claims by test developers or test users that a test, manual, or procedure satisfies or follows these standards should be made with

care. It is appropriate for developers or users to state that efforts were made to adhere to the *Standards*, and to provide documents describing and supporting those efforts. Blanket claims without supporting evidence should not be made.

4) These standards are concerned with a field that is evolving. Consequently, there is a continuing need to monitor changes in the field and to revise this document as knowledge develops.

5) Prescription of the use of specific technical methods is not the intent of the *Standards*. For example, where specific statistical reporting requirements are mentioned, the phrase "or generally accepted equivalent" always should be understood.

The standards do not attempt to repeat or to incorporate the many legal or regulatory requirements that might be relevant to the issues they address. In some areas, such as the collection, analysis, and use of test data and results for different subgroups, the law may both require participants in the testing process to take certain actions and prohibit those participants from taking other actions. Where it is apparent that one or more standards or comments address an issue on which established legal requirements may be particularly relevant, the standard, comment, or introductory material may make note of that fact. Lack of specific reference to legal requirements, however, does not imply that no relevant requirement exists. In all situations, participants in the testing process should separately consider and, where appropriate, obtain legal advice on legal and regulatory requirements.

The Number of Standards

The number of standards has increased from the 1985 *Standards* for a variety of reasons. First, and most importantly, new developments have led to the addition of new standards. Commonly these deal with new types

INTRODUCTION

of tests or new uses for existing tests, rather than being broad standards applicable to all tests. Second, on the basis of recognition that some users of the *Standards* may turn only to chapters directly relevant to a given application, certain standards are repeated in different chapters. When such repetition occurs, the essence of the standard is the same. Only the wording, area of application, or elaboration in the comment is changed. Third, standards dealing with important nontechnical issues, such as avoiding conflicts of interest and equitable treatment of all test takers, have been added. Although such topics have not been addressed in prior versions of the *Standards*, they are not likely to be viewed as imposing burdensome new requirements. Thus the increase in the number of standards does not per se signal an increase in the obligations placed on test developers and test users.

Tests as Measures of Constructs

We depart from some historical uses of the term "construct," which reserve the term for characteristics that are not directly observable, but which are inferred from interrelated sets of observations. This historical perspective invites confusion. Some tests are viewed as measures of constructs, while others are not. In addition, considerable debate has ensued as to whether certain characteristics measured by tests are properly viewed as constructs. Furthermore, the types of validity evidence thought to be suitable can differ as a result of whether a given test is viewed as measuring a construct.

We use the term *construct* more broadly as the concept or characteristic that a test is designed to measure. Rarely, if ever, is there a single possible meaning that can be attached to a test score or a pattern of test responses. Thus, it is always incumbent on a testing professional to specify the construct interpretation that will be made on the basis of the

score or response pattern. The notion that some tests are not under the purview of the *Standards* because they do not measure constructs is contrary to this use of the term. Also, as detailed in chapter 1, evolving conceptualizations of the concept of validity no longer speak of different types of validity but speak instead of different lines of validity evidence, all in service of providing information relevant to a specific intended interpretation of test scores. Thus, many lines of evidence can contribute to an understanding of the construct meaning of test scores.

Organization of This Volume

Part I of the *Standards*, "Test Construction, Evaluation, and Documentation," contains standards for validity (ch. 1); reliability and errors of measurement (ch. 2); test development and revision (ch. 3); scaling, norming, and score comparability (ch. 4); test administration, scoring, and reporting (ch. 5); and supporting documentation for tests (ch. 6). Part II addresses "Fairness in Testing," and contains standards on fairness and bias (ch. 7); the rights and responsibilities of test takers (ch. 8); testing individuals of diverse linguistic backgrounds (ch. 9); and testing individuals with disabilities (ch. 10). Part III treats specific "Testing Applications," and contains standards involving general responsibilities of test users (ch. 11); psychological testing and assessment (ch. 12); educational testing and assessment (ch. 13); testing in employment and credentialing (ch. 14); and testing in program evaluation and public policy (ch. 15).

Each chapter begins with introductory text that provides background for the standards that follow. This revision of the *Standards* contains more extensive introductory text material than its predecessor. Recognizing the common use of the *Standards* in the education of future test developers and users, the committee opted to provide a context for the standards themselves by pre-

INTRODUCTION

senting more background material than in previous versions. This text is designed to assist in the interpretation of the standards that follow in each chapter. Although the text is at times prescriptive and exhortatory, it should not be interpreted as imposing additional standards.

The *Standards* also contains an index and includes a glossary that provides definitions for terms as they are specifically used in this volume.

PART I

**Test Construction,
Evaluation, and
Documentation**

1. VALIDITY

Background

Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. When test scores are used or interpreted in more than one way, each intended interpretation must be validated.

Validation logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use. The proposed interpretation refers to the construct or concepts the test is intended to measure. Examples of constructs are mathematics achievement, performance as a computer technician, depression, and self-esteem. To support test development, the proposed interpretation is elaborated by describing its scope and extent and by delineating the aspects of the construct that are to be represented. The detailed description provides a conceptual framework for the test, delineating the knowledge, skills, abilities, processes, or characteristics to be assessed. The framework indicates how this representation of the construct is to be distinguished from other constructs and how it should relate to other variables.

The conceptual framework is partially shaped by the ways in which test scores will be used. For instance, a test of mathematics achievement might be used to place a student in an appropriate program of instruction, to endorse a high school diploma, or to inform a college admissions decision. Each of these uses implies a somewhat different interpretation of the mathematics achievement test

scores: that a student will benefit from a particular instructional intervention, that a student has mastered a specified curriculum, or that a student is likely to be successful with college-level work. Similarly, a test of self-esteem might be used for psychological counseling, to inform a decision about employment, or for the basic scientific purpose of elaborating the construct of self-esteem. Each of these potential uses shapes the specified framework and the proposed interpretation of the test's scores and also has implications for test development and evaluation.

Validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use. The conceptual framework points to the kinds of evidence that might be collected to evaluate the proposed interpretation in light of the purposes of testing. As validation proceeds, and new evidence about the meaning of a test's scores becomes available, revisions may be needed in the test, in the conceptual framework that shapes it, and even in the construct underlying the test.

The wide variety of tests and circumstances makes it natural that some types of evidence will be especially critical in a given case, whereas other types will be less useful. The decision about what types of evidence are important for validation in each instance can be clarified by developing a set of propositions that support the proposed interpretation for the particular purpose of testing. For instance, when a mathematics achievement test is used to assess readiness for an advanced course, evidence for the following propositions might be deemed necessary: (a) that certain skills are prerequisite for the advanced course; (b) that the content domain of the test is consistent with these prerequisite skills; (c) that test scores can be generalized across relevant sets of items; (d) that test scores are not unduly influenced by ancillary variables,

such as writing ability; (e) that success in the advanced course can be validly assessed; and (f) that examinees with high scores on the test will be more successful in the advanced course than examinees with low scores on the test. Examples of propositions in other testing contexts might include, for instance, the proposition that examinees with high general anxiety scores experience significant anxiety in a range of settings, the proposition that a child's score on an intelligence scale is strongly related to the child's academic performance, or the proposition that a certain pattern of scores on a neuropsychological battery indicates impairment characteristic of brain injury. The validation process evolves as these propositions are articulated and evidence is gathered to evaluate their soundness.

Identifying the propositions implied by a proposed test interpretation can be facilitated by considering rival hypotheses that may challenge the proposed interpretation. It is also useful to consider the perspectives of different interested parties, existing experience with similar tests and contexts, and the expected consequences of the proposed test use. Plausible rival hypotheses can often be generated by considering whether a test measures less or more than its proposed construct. Such concerns are referred to as *construct underrepresentation* and *construct-irrelevant variance*.

Construct underrepresentation refers to the degree to which a test fails to capture important aspects of the construct. It implies a narrowed meaning of test scores because the test does not adequately sample some types of content, engage some psychological processes, or elicit some ways of responding that are encompassed by the intended construct. Take, for example, a test of reading comprehension intended to measure children's ability to read and interpret stories with understanding. A particular test might underrepresent the intended construct because it did not contain a sufficient variety of read-

ing passages or ignored a common type of reading material. As another example, a test of anxiety might measure only physiological reactions and not emotional, cognitive, or situational components.

Construct-irrelevant variance refers to the degree to which test scores are affected by processes that are extraneous to its intended construct. The test scores may be systematically influenced to some extent by components that are not part of the construct. In the case of a reading comprehension test, construct-irrelevant components might include an emotional reaction to the test content, familiarity with the subject matter of the reading passages on the test, or the writing skill needed to compose a response. Depending on the detailed definition of the construct, vocabulary knowledge or reading speed might also be irrelevant components. On a test of anxiety, a response bias to underreport anxiety might be considered a source of construct-irrelevant variance.

Nearly all tests leave out elements that some potential users believe should be measured and include some elements that some potential users consider inappropriate. Validation involves careful attention to possible distortions in meaning arising from inadequate representation of the construct and also to aspects of measurement such as test format, administration conditions, or language level that may materially limit or qualify the interpretation of test scores. That is, the process of validation may lead to revisions in the test, the conceptual framework of the test, or both. The revised test would then need validation.

When propositions have been identified that would support the proposed interpretation of test scores, validation can proceed by developing empirical evidence, examining relevant literature, and/or conducting logical analyses to evaluate each of these propositions. Empirical evidence may include both local evidence, produced within the contexts where the test will be used, and evidence from similar testing

applications in other settings. Use of existing evidence from similar tests and contexts can enhance the quality of the validity argument, especially when current data are limited.

Because a validity argument typically depends on more than one proposition, strong evidence in support of one in no way diminishes the need for evidence to support others. For example, a strong predictor-criterion relationship in an employment setting is not sufficient to justify test use for selection without considering the appropriateness and meaningfulness of the criterion measure. Professional judgment guides decisions regarding the specific forms of evidence that can best support the intended interpretation and use. As in all scientific endeavors, the quality of the evidence is primary. A few lines of solid evidence regarding a particular proposition are better than numerous lines of evidence of questionable quality.

Validation is the joint responsibility of the test developer and the test user. The test developer is responsible for furnishing relevant evidence and a rationale in support of the intended test use. The test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used. When the use of a test differs from that supported by the test developer, the test user bears special responsibility for validation. The standards apply to the validation process, for which the appropriate parties share responsibility. It should be noted that important contributions to the validity evidence are made as other researchers report findings of investigations that are related to the meaning of scores on the test.

Sources of Validity Evidence

The following sections outline various sources of evidence that might be used in evaluating a proposed interpretation of test scores for particular purposes. These sources of evidence may illuminate different aspects of validity,

but they do not represent distinct types of validity. Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose. Like the 1985 *Standards*, this edition refers to types of validity evidence, rather than distinct types of validity. To emphasize this distinction, the treatment that follows does not follow traditional nomenclature (i.e., the use of the terms *content validity* or *predictive validity*). The glossary contains definitions of the traditional terms, explicating the difference between traditional and current use.

EVIDENCE BASED ON TEST CONTENT

Important validity evidence can be obtained from an analysis of the relationship between a test's content and the construct it is intended to measure. Test content refers to the themes, wording, and format of the items, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring. Test developers often work from a specification of the content domain. The content specification carefully describes the content in detail, often with a classification of areas of content and types of items. Evidence based on test content can include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores. Evidence based on content can also come from expert judgments of the relationship between parts of the test and the construct. For example, in developing a licensure test, the major facets of the specific occupation can be specified, and experts in that occupation can be asked to assign test items to the categories defined by those facets. They, or other qualified experts, can then judge the representativeness of the chosen set of items. Sometimes rules or algorithms can be constructed to select or generate items that differ systematically on the various facets of content, according to specifications.

Some tests are based on systematic observations of behavior. For example, a listing of the tasks comprising a job domain may be developed from observations of behavior in a job, together with judgments of subject-matter experts. Expert judgments can be used to assess the relative importance, criticality, and/or frequency of the various tasks. A job sample test can then be constructed from a random or stratified sampling of tasks rated highly on these characteristics. The test can then be administered under standardized conditions in an off-the-job setting.

The appropriateness of a given content domain is related to the specific inferences to be made from test scores. Thus, when considering an available test for a purpose other than that for which it was first developed, it is especially important to evaluate the appropriateness of the original content domain for the proposed new use. In educational program evaluations, for example, tests may properly cover material that receives little or no attention in the curriculum, as well as that toward which instruction is directed. Policymakers can then evaluate student achievement with respect to both content neglected and content addressed. On the other hand, when student mastery of a delivered curriculum is tested for purposes of informing decisions about individual students, such as promotion or graduation, the framework elaborating a content domain is appropriately limited to what students have had an opportunity to learn from the curriculum as delivered.

Evidence about content can be used, in part, to address questions about differences in the meaning or interpretation of test scores across relevant subgroups of examinees. Of particular concern is the extent to which construct underrepresentation or construct-irrelevant components may give an unfair advantage or disadvantage to one or more subgroups of examinees. Careful review of the construct and test content domain by a diverse panel of experts may point to potential sources of

irrelevant difficulty (or easiness) that require further investigation.

EVIDENCE BASED ON RESPONSE PROCESSES

Theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees. For instance, if a test is intended to assess mathematical reasoning, it becomes important to determine whether examinees are, in fact, reasoning about the material given instead of following a standard algorithm. For another instance, scores on a scale intended to assess the degree of an individual's extroversion or introversion should not be strongly influenced by social conformity.

Evidence based on response processes generally comes from analyses of individual responses. Questioning test takers about their performance strategies or responses to particular items can yield evidence that enriches the definition of a construct. Maintaining records that monitor the development of a response to a writing task, through successive written drafts or electronically monitored revisions, for instance, also provides evidence of process. Documentation of other aspects of performance, like eye movements or response times, may also be relevant to some constructs. Inferences about processes involved in performance can also be developed by analyzing the relationship among parts of the test and between the test and other variables. Wide individual differences in process can be revealing and may lead to reconsideration of certain test formats.

Evidence of response processes can contribute to questions about differences in meaning or interpretation of test scores across relevant subgroups of examinees. Process studies involving examinees from different subgroups can assist in determining the extent to which capabilities irrelevant or ancillary to the construct may be differentially influencing their performance.

Studies of response processes are not limited to the examinee. Assessments often rely on observers or judges to record and/or evaluate examinees' performances or products. In such cases, relevant validity evidence includes the extent to which the processes of observers or judges are consistent with the intended interpretation of scores. For instance, if judges are expected to apply particular criteria in scoring examinees' performances, it is important to ascertain whether they are, in fact, applying the appropriate criteria and not being influenced by factors that are irrelevant to the intended interpretation. Thus, validation may include empirical studies of how observers or judges record and evaluate data along with analyses of the appropriateness of these processes to the intended interpretation or construct definition.

EVIDENCE BASED ON INTERNAL STRUCTURE

Analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based. The conceptual framework for a test may imply a single dimension of behavior, or it may posit several components that are each expected to be homogeneous, but that are also distinct from each other. For example, a measure of discomfort on a health survey might assess both physical and emotional health. The extent to which item interrelationships bear out the presumptions of the framework would be relevant to validity.

The specific types of analysis and their interpretation depend on how the test will be used. For example, if a particular application posited a series of test components of increasing difficulty, empirical evidence of the extent to which response patterns conformed to this expectation would be provided. A theory that posited unidimensionality would call for evidence of item homogeneity. In this case, the item interrelationships

also provide an estimate of score reliability, but such an index would be inappropriate for tests with a more complex internal structure.

Some studies of the internal structure of tests are designed to show whether particular items may function differently for identifiable subgroups of examinees. Differential item functioning occurs when different groups of examinees with similar overall ability, or similar status on an appropriate criterion, have, on average, systematically different responses to a particular item. This issue is discussed in chapters 3 and 7. However, differential item functioning is not always a flaw or weakness. Subsets of items that have a specific characteristic in common (e.g., specific content, task representation) may function differently for different groups of similarly scoring examinees. This indicates a kind of multidimensionality that may be unexpected or may conform to the test framework.

EVIDENCE BASED ON RELATIONS TO OTHER VARIABLES

Analyses of the relationship of test scores to variables external to the test provide another important source of validity evidence. External variables may include measures of some criteria that the test is expected to predict, as well as relationships to other tests hypothesized to measure the same constructs, and tests measuring related or different constructs. Measures other than test scores, such as performance criteria, are often used in employment settings. Categorical variables, including group membership variables, become relevant when the theory underlying a proposed test use suggests that group differences should be present or absent if a proposed test interpretation is to be supported. Evidence based on relationships with other variables addresses questions about the degree to which these relationships are consistent with the construct underlying the proposed test interpretations.

Convergent and discriminant evidence. Relationships between test scores and other measures intended to assess similar constructs provide convergent evidence, whereas relationships between test scores and measures purportedly of different constructs provide discriminant evidence. For instance, within some theoretical frameworks, scores on a multiple-choice test of reading comprehension might be expected to relate closely (convergent evidence) to other measures of reading comprehension based on other methods, such as essay responses; conversely, test scores might be expected to relate less closely (discriminant evidence) to measures of other skills, such as logical reasoning. Relationships among different methods of measuring the construct can be especially helpful in sharpening and elaborating score meaning and interpretation.

Evidence of relations with other variables can involve experimental as well as correlational evidence. Studies might be designed, for instance, to investigate whether scores on a measure of anxiety improve as a result of some psychological treatment or whether scores on a test of academic achievement differentiate between instructed and noninstructed groups. If performance increases due to short-term coaching are viewed as a threat to validity, it would be useful to investigate whether coached and uncoached groups perform differently.

Test-criterion relationships. Evidence of the relation of test scores to a relevant criterion may be expressed in various ways, but the fundamental question is always: How accurately do test scores predict criterion performance? The degree of accuracy deemed necessary depends on the purpose for which the test is used.

The criterion variable is a measure of some attribute or outcome that is of primary interest, as determined by test users, who may be administrators in a school system, the management of a firm, or clients. The choice of

the criterion and the measurement procedures used to obtain criterion scores are of central importance. The value of a test-criterion study depends on the relevance, reliability, and validity of the interpretation based on the criterion measure for a given testing application.

Historically, two designs, often called predictive and concurrent, have been distinguished for evaluating test-criterion relationships. A predictive study indicates how accurately test data can predict criterion scores that are obtained at a later time. A concurrent study obtains predictor and criterion information at about the same time. When prediction is actually contemplated, as in education or employment settings, or in planning rehabilitation regimens, predictive studies can retain the temporal differences and other characteristics of the practical situation. Concurrent evidence, which avoids temporal changes, is particularly useful for psychodiagnostic tests or to investigate alternative measures of some specified construct. In general, the choice of research strategy is guided by prior evidence of the extent to which predictive and concurrent studies yield the same or different results in the domain.

Test scores are sometimes used in allocating individuals to different treatments, such as different jobs within an institution, in a way that is advantageous for the institution and for the individuals. In that context, evidence is needed to judge the suitability of using a test when classifying or assigning a person to one job versus another or to one treatment versus another. Classification decisions are supported by evidence that the relationship of test scores to performance criteria is different for different treatments. It is possible for tests to be highly predictive of performance for different education programs or jobs without providing the information necessary to make a comparative judgment of the efficacy of assignments or treatments. In general, decision rules for selection or placement are also influenced by the number of persons to be accepted or the

numbers that can be accommodated in alternative placement categories.

Evidence about relations to other variables is also used to investigate questions of differential prediction for groups. For instance, a finding that the relation of test scores to a relevant criterion variable differs from one group to another may imply that the meaning of the scores is not the same for members of the different groups, perhaps due to construct underrepresentation or construct-irrelevant components. However, the difference may also imply that the criterion has different meaning for different groups. The differences in test-criterion relationships can also arise from measurement error, especially when group means differ, so such differences do not necessarily indicate differences in score meaning. (See chapter 7.)

Validity generalization. An important issue in educational and employment settings is the degree to which evidence of validity based on test-criterion relations can be generalized to a new situation without further study of validity in that new situation. When a test is used to predict the same or similar criteria (e.g., performance of a given job) at different times or in different places, it is typically found that observed test-criterion correlations vary substantially. In the past, this has been taken to imply that local validation studies are always required. More recently, meta-analytic analyses have shown that in some domains, much of this variability may be due to statistical artifacts such as sampling fluctuations and variations across validation studies in the ranges of test scores and in the reliability of criterion measures. When these and other influences are taken into account, it may be found that the remaining variability in validity coefficients is relatively small. Thus, statistical summaries of past validation studies in similar situations may be useful in estimating test-criterion relationships in a new situation. This practice is referred to as the study of validity generalization.

In some circumstances, there is a strong basis for using validity generalization. This would be the case where the meta-analytic database is large, where the meta-analytic data adequately represent the type of situation to which one wishes to generalize, and where correction for statistical artifacts produces a clear and consistent pattern of validity evidence. In such circumstances, the informational value of a local validity study may be relatively limited. In other circumstances, the inferential leap required for generalization may be much larger. The meta-analytic database may be small, the findings may be less consistent, or the new situation may involve features markedly different from those represented in the meta-analytic database. In such circumstances, situation-specific evidence of validity will be relatively more informative. Although research on validity generalization shows that results of a single local validation study may be quite imprecise, there are situations where a single study, carefully done, with adequate sample size, provides sufficient evidence to support test use in a new situation. This highlights the importance of examining carefully the comparative informational value of local versus meta-analytic studies.

In conducting studies of the generalizability of validity evidence, the prior studies that are included may vary according to several situational facets. Some of the major facets are (a) differences in the way the predictor construct is measured, (b) the type of job or curriculum involved, (c) the type of criterion measure used, (d) the type of test takers, and (e) the time period in which the study was conducted. In any particular study of validity generalization, any number of these facets might vary, and a major objective of the study is to determine empirically the extent to which variation in these facets affects the test-criterion correlations obtained.

The extent to which predictive or concurrent evidence of validity generalization can

be used in new situations is in large measure a function of accumulated research. Although *evidence of generalization can often help to support a claim of validity in a new situation, the extent of available data limits the extent to which the claim can be sustained.*

The above discussion focuses on the use of cumulative databases to estimate predictor-criterion relationships. Meta-analytic techniques can also be used to summarize other forms of data relevant to other inferences one may wish to draw from test scores in a particular application, such as effects of coaching and effects of certain alterations in testing conditions to accommodate test takers with certain disabilities.

EVIDENCE BASED ON CONSEQUENCES OF TESTING

An issue receiving attention in recent years is the incorporation of the intended and unintended consequences of test use into the concept of validity. *Evidence about consequences can inform validity decisions. Here, however, it is important to distinguish between evidence that is directly relevant to validity and evidence that may inform decisions about social policy but falls outside the realm of validity.*

Distinguishing between issues of validity and issues of social policy becomes particularly important in cases where differential consequences of test use are observed for different identifiable groups. For example, concerns have been raised about the effect of group differences in test scores on employment selection and promotion, the placement of children in special education classes, and the narrowing of a school's curriculum to exclude learning of objectives that are not assessed. Although information about the consequences of testing may influence decisions about test use, such consequences do not in and of themselves detract from the validity of intended test interpretations. Rather, judgments of validity or invalidity in the light of testing

consequences depend on a more searching inquiry into the sources of those consequences.

Take, as an example, a *finding of different hiring rates for members of different groups as a consequence of using an employment test. If the difference is due solely to an unequal distribution of the skills the test purports to measure, and if those skills are, in fact, important contributors to job performance, then the finding of group differences per se does not imply any lack of validity for the intended inference. If, however, the test measured skill differences unrelated to job performance (e.g., a sophisticated reading test for a job that required only minimal functional literacy), or if the differences were due to the test's sensitivity to some examinee characteristic not intended to be part of the test construct, then validity would be called into question, even if test scores correlated positively with some measure of job performance. Thus, evidence about consequences may be directly relevant to validity when it can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant components. Evidence about consequences that cannot be so traced—that in fact reflects valid differences in performance—is crucial in informing policy decisions but falls outside the technical purview of validity.*

Tests are commonly administered in the expectation that some benefit will be realized from the intended use of the scores. A few of the many possible benefits are selection of efficacious treatments for therapy, placement of workers in suitable jobs, prevention of unqualified individuals from entering a profession, or improvement of classroom instructional practices. A fundamental purpose of validation is to indicate whether these specific benefits are likely to be realized. Thus, in the case of a test used in placement decisions, the validation would be informed by evidence that alternative placements, in fact, are differentially beneficial to the persons and the institution. In the case of employment testing,

if a test publisher claims that use of the test will result in reduced employee training costs, improved workforce efficiency, or some other benefit, then the validation would be informed by evidence in support of that claim.

Claims are sometimes made for benefits of testing that go beyond direct uses of the test scores themselves. Educational tests, for example, may be advocated on the grounds that their use will improve student motivation or encourage changes in classroom instructional practices by holding educators accountable for valued learning outcomes. Where such claims are central to the rationale advanced for testing, the direct examination of testing consequences necessarily assumes even greater importance. The validation process in such cases would be informed by evidence that the anticipated benefits of testing are being realized.

Integrating the Validity Evidence

A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. It encompasses evidence gathered from new studies and evidence available from earlier reported research. The validity argument may indicate the need for refining the definition of the construct, may suggest revisions in the test or other aspects of the testing process, and may indicate areas needing further study.

Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. This includes evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all examinees, as described in subsequent chapters of the *Standards*.

Standard 1.1

A rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation.

Comment: The rationale should indicate what propositions are necessary to investigate the intended interpretation. The comprehensive summary should combine logical analysis with empirical evidence to provide support for the test rationale. Evidence may come from studies conducted locally, in the setting where the test is to be used; from specific prior studies; or from comprehensive statistical syntheses of available studies meeting clearly specified criteria. No type of evidence is inherently preferable to others; rather, the quality and relevance of the evidence to the intended test use determine the value of a particular kind of evidence. A presentation of empirical evidence on any point should give due weight to all relevant findings in the scientific literature, including those inconsistent with the intended interpretation or use. Test developers have the responsibility to provide support for their own recommendations, but test users are responsible for evaluating the quality of the validity evidence provided and its relevance to the local situation.

Standard 1.2

The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, and the construct that the test is intended to assess should be clearly described.

Comment: Statements about validity should refer to particular interpretations and uses. It is incorrect to use the unqualified phrase “the validity of the test.” No test is valid for all purposes or in all situations. Each recom-

STANDARDS

VALIDITY / PART I

mended use or interpretation requires validation and should specify in clear language the population for which the test is intended, the construct it is intended to measure, and the manner and contexts in which test scores are to be employed.

Standard 1.3

If validity for some common or likely interpretation has not been investigated, or if the interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be cautioned about making unsupported interpretations.

Comment: If past experience suggests that a test is likely to be used inappropriately for certain kinds of decisions, specific warnings against such uses should be given. On the other hand, no two situations are ever identical, so some generalization by the user is always necessary. Professional judgment is required to evaluate the extent to which existing validity evidence supports a given test use.

Standard 1.4

If a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary.

Comment: Professional judgment is required to evaluate the extent to which existing validity evidence applies in the new situation and to determine what new evidence may be needed. The amount and kinds of new evidence required may be influenced by experience with similar prior test uses or interpretations and by the amount, quality, and relevance of existing data.

Standard 1.5

The composition of any sample of examinees from which validity evidence is

obtained should be described in as much detail as is practical, including major relevant sociodemographic and developmental characteristics.

Comment: Statistical findings can be influenced by factors affecting the sample on which the results are based. When the sample is intended to represent a population, that population should be described, and attention should be drawn to any systematic factors that may limit the representativeness of the sample. Factors that might reasonably be expected to affect the results include self-selection, attrition, linguistic prowess, disability status, and exclusion criteria, and others. If the subjects of a validity study are patients, for example, then the diagnoses of the patients are important, as well as other characteristics, such as the severity of the diagnosed condition. For tests used in industry, the employment status (e.g., applicants versus current job holders), the general level of experience and educational background and the gender and ethnic composition of the sample may be relevant information. For tests used in educational settings, relevant information may include educational background, developmental level, community characteristics, or school admissions policies, as well as the gender and ethnic composition of the sample. Sometimes restrictions about privacy preclude obtaining such population information.

Standard 1.6

When the validation rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified in reference to the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified.

Comment: For example, test developers might provide a logical structure that maps the items on the test to the content domain, illustrating the relevance of each item and the adequacy with which the set of items represents the content domain. Areas of the content domain that are not included among the test items could be indicated as well.

Standard 1.7

When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications, and experience, of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth.

Comment: Systematic collection of judgments or opinions may occur at many points in test construction (e.g., in eliciting expert judgments of content appropriateness or adequate content representation), in formulating rules or standards for score interpretation (e.g., in setting cut scores), or in test scoring (e.g., rating of essay responses). Whenever such procedures are employed, the quality of the resulting judgments is important to the validation. It may be entirely appropriate to have experts work together to reach consensus, but it would not then be appropriate to treat their respective judgments as statistically independent.

Standard 1.8

If the rationale for a test use or score interpretation depends on premises about the psychological processes or cognitive opera-

tions used by examinees, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by observers or scorers are part of the argument for validity, similar information should be provided.

Comment: If the test specification delineates the processes to be assessed, then evidence is needed that the test items do, in fact, tap the intended processes.

Standard 1.9

If a test is claimed to be essentially unaffected by practice and coaching, then the sensitivity of test performance to change with these forms of instruction should be documented.

Comment: Materials to aid in score interpretation should summarize evidence indicating the degree to which improvement with practice or coaching can be expected. Also, materials written for test takers should provide practical guidance about the value of test preparation activities, including coaching.

Standard 1.10

When interpretation of performance on specific items, or small subsets of items, is suggested, the rationale and relevant evidence in support of such interpretation should be provided. When interpretation of individual item responses is likely but is not recommended by the developer, the user should be warned against making such interpretations.

Comment: Users should be given sufficient guidance to enable them to judge the degree of confidence warranted for any use or interpretation recommended by the test developer. Test manuals and score reports should discourage overinterpretation of information that may be subject to considerable error. This is especially important if interpretation

STANDARDS

VALIDITY / PART I

of performance on isolated items, small subsets of items, or subtest scores is suggested.

Standard 1.11

If the rationale for a test use or interpretation depends on premises about the relationships among parts of the test, evidence concerning the internal structure of the test should be provided.

Comment: It might be claimed, for example, that a test is essentially unidimensional. Such a claim could be supported by a multivariate statistical analysis, such as a factor analysis, showing that the score variability attributable to one major dimension was much greater than the score variability attributable to any other identified dimension. When a test provides more than one score, the interrelationships of those scores should be shown to be consistent with the construct(s) being assessed.

Standard 1.12

When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given.

Comment: When a test provides more than one score, the distinctiveness of the separate scores should be demonstrated, and the interrelationships of those scores should be shown to be consistent with the construct(s) being assessed. Moreover, evidence for the validity of interpretations of two separate scores would not necessarily justify an interpretation of the difference between them. Rather, the rationale and supporting evidence must pertain directly to the specific score or score combination to be interpreted or used.

Standard 1.13

When validity evidence includes statistical analyses of test results, either alone or together with data on other variables, the conditions under which the data were collected should be described in enough detail that users can judge the relevance of the statistical findings to local conditions. Attention should be drawn to any features of a validation data collection that are likely to differ from typical operational testing conditions and that could plausibly influence test performance.

Comment: Such conditions might include (but would not be limited to) the following: examinee motivation or prior preparation, the distribution of test scores over examinees, the time allowed for examinees to respond or other administrative conditions, examiner training or other examiner characteristics, the time intervals separating collection of data on different measures, or conditions that may have changed since the validity evidence was obtained.

Standard 1.14

When validity evidence includes empirical analyses of test responses together with data on other variables, the rationale for selecting the additional variables should be provided. Where appropriate and feasible, evidence concerning the constructs represented by other variables, as well as their technical properties, should be presented or cited. Attention should be drawn to any likely sources of dependence (or lack of independence) among variables other than dependencies among the construct(s) they represent.

Comment: The patterns of association between and among scores on the instrument under study and other variables should be consistent with theoretical expectations. The additional variables might be demographic

STANDARDS

characteristics, indicators of treatment conditions, or scores on other measures. They might include intended measures of the same construct or of different constructs. The reliability of scores from such other measures and the validity of intended interpretations of scores from these measures are an important part of the validity evidence for the instrument under study. If such variables include composite scores, the construction of the composites should be explained. In addition to considering the properties of each variable in isolation, it is important to guard against faulty interpretations arising from spurious sources of dependency among measures, including correlated errors or shared variance due to common methods of measurement or common elements.

Standard 1.15

When it is asserted that a certain level of test performance predicts adequate or inadequate criterion performance, information about the levels of criterion performance associated with given levels of test scores should be provided.

Comment: Regression equations are more useful than correlation coefficients, which are generally insufficient to fully describe patterns of association between tests and other variables. Means, standard deviations, and other statistical summaries are needed, as well as information about the distribution of criterion performances conditional upon a given test score. Evidence of overall association between variables should be supplemented by information about the form of that association and about the variability associated with that association in different ranges of test scores. Note that data collections employing examinees selected for their extreme scores on one or more measures (extreme groups) typically cannot provide adequate information about the association.

Standard 1.16

When validation relies on evidence that test scores are related to one or more criterion variables, information about the suitability and technical quality of the criteria should be reported.

Comment: The description of each criterion variable should include evidence concerning its reliability, the extent to which it represents the intended construct (e.g., job performance), and the extent to which it is likely to be influenced by extraneous sources of variance. Special attention should be given to sources that previous research suggests may introduce extraneous variance that might bias the criterion for or against identifiable groups.

Standard 1.17

If test scores are used in conjunction with other quantifiable variables to predict some outcome or criterion, regression (or equivalent) analyses should include those additional relevant variables along with the test scores.

Comment: In general, if several predictors of some criterion are available, the optimum combination of predictors cannot be determined solely from separate, pairwise examinations of the criterion variable with each separate predictor in turn. It is often informative to estimate the increment in predictive accuracy that may be expected when each variable, including the test score, is introduced in addition to all other available variables. Analyses involving multiple predictors should be verified by cross-validation or equivalent analysis whenever feasible, and the precision of estimated regression coefficients should be reported.

Standard 1.18

When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coeffi-

STANDARDS

VALIDITY / PART I

cients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported.

Comment: The correlation between two variables, such as test scores and criterion measures, depends on the range of values on each variable. For example, the test scores and the criterion values of selected applicants will typically have a smaller range than the scores of all applicants. Statistical methods are available for adjusting the correlation to reflect the population of interest rather than the sample available. Such adjustments are often appropriate, as when comparing results across various situations. Reporting an adjusted correlation should be accompanied by a statement of the method and the statistics used in making the adjustment.

Standard 1.19

If a test is recommended for use in assigning persons to alternative treatments or is likely to be so used, and if outcomes from those treatments can reasonably be compared on a common criterion, then, whenever feasible, supporting evidence of differential outcomes should be provided.

Comment: If a test is used for classification into alternative occupational, therapeutic, or educational programs, it is not sufficient just to show that the test predicts treatment outcomes. Support for the validity of the classification procedure is provided by showing that the test is useful in determining which persons are likely to profit differentially from one treatment or another. Treatment categories may have to be combined to assemble sufficient cases for statistical analysis. It is recognized, however, that such research may not be feasible, because ethical and legal constraints on differential assignments may forbid control groups.

Standard 1.20

When a meta-analysis is used as evidence of the strength of a test-criterion relationship, the test and the criterion variables in the local situation should be comparable with those in the studies summarized. If relevant research includes credible evidence that any other features of the testing application may influence the strength of the test-criterion relationship, the correspondence between those features in the local situation and in the meta-analysis should be reported. Any significant disparities that might limit the applicability of the meta-analytic findings to the local situation should be noted explicitly.

Comment: The meta-analysis should incorporate all available studies meeting explicitly stated inclusion criteria. Meta-analytic evidence used in test validation typically is based on a number of tests measuring the same or very similar constructs and criterion measures that likewise measure the same or similar constructs. A meta-analytic study may also be limited to a single test and a single criterion. For each study included in the analysis, the test-criterion relationship is expressed in some common metric, often as an *effect size*. The strength of the test-criterion relationship may be moderated by features of the situation in which the test and criterion measures were obtained (e.g., types of jobs, characteristics of test takers, time interval separating collection of test and criterion measures, year or decade in which the data were collected). If test-criterion relationships vary according to such moderator variables, then, the numbers of studies permitting, the meta-analysis should report separate estimated effect size distributions conditional upon relevant situational features. This might be accomplished, for example, by reporting separate distributions for subsets of studies or by estimating the magnitudes of the influences of situational features on effect sizes.

Standard 1.21

Any meta-analytic evidence used to support an intended test use should be clearly described, including methodological choices in identifying and coding studies, correcting for artifacts, and examining potential moderator variables. Assumptions made in correcting for artifacts such as criterion unreliability and range restriction should be presented, and the consequences of these assumptions made clear.

Comment: Meta-analysis inevitably involves judgments regarding a number of methodological choices. The bases for these judgments should be articulated. In the case of choices involving some degree of uncertainty, such as artifact corrections based on assumed values, the uncertainty should be acknowledged and the degree to which conclusions about validity hinge on these assumptions should be examined and reported.

Standard 1.22

When it is clearly stated or implied that a recommended test use will result in a specific outcome, the basis for expecting that outcome should be presented, together with relevant evidence.

Comment: If it is asserted, for example, that using a given test for employee selection will result in reduced employee errors or training costs, evidence in support of that assertion should be provided. A given claim for the benefits of test use may be supported by logical or theoretical argument as well as empirical data. Due weight should be given to findings in the scientific literature that may be inconsistent with the stated expectation.

Standard 1.23

When a test use or score interpretation is recommended on the grounds that testing or

the testing program per se will result in some indirect benefit in addition to the utility of information from the test scores themselves, the rationale for anticipating the indirect benefit should be made explicit. Logical or theoretical arguments and empirical evidence for the indirect benefit should be provided. Due weight should be given to any contradictory findings in the scientific literature, including findings suggesting important indirect outcomes other than those predicted.

Comment: For example, certain educational testing programs have been advocated on the grounds that they would have a salutary influence on classroom instructional practices or would clarify students' understanding of the kind or level of achievement they were expected to attain. To the extent that such claims enter into the justification for a testing program, they become part of the validity argument for test use and so should be examined as part of the validation effort. Due weight should be given to evidence against such predictions, for example, evidence that under some conditions educational testing may have a negative effect on classroom instruction.

Standard 1.24

When unintended consequences result from test use, an attempt should be made to investigate whether such consequences arise from the test's sensitivity to characteristics other than those it is intended to assess or to the test's failure fully to represent the intended construct.

Comment: The validity of test score interpretations may be limited by construct-irrelevant components or construct underrepresentation. When unintended consequences appear to stem, at least in part, from the use of one or more tests, it is especially important to check

STANDARDS

that these consequences do not arise from such sources of invalidity. Although group differences, in and of themselves, do not call into question the validity of a proposed interpretation, they may increase the salience of plausible rival hypotheses that should be investigated as part of the validation effort.

2. RELIABILITY AND ERRORS OF MEASUREMENT

Background

A test, broadly defined, is a set of tasks designed to elicit or a scale to describe examinee behavior in a specified domain, or a system for collecting samples of an individual's work in a particular area. Coupled with the device is a scoring procedure that enables the examiner to quantify, evaluate, and interpret the behavior or work samples. *Reliability* refers to the consistency of such measurements when the testing procedure is repeated on a population of individuals or groups.

The discussion that follows introduces concepts and procedures that may not be familiar to some readers. It is not expected that the brief definitions and explanations presented here will be sufficient to enable the less sophisticated reader to become adequately conversant with these developments. To achieve a better understanding, such readers may need to consult more comprehensive treatments in the measurement literature.

The usefulness of behavioral measurements presupposes that individuals and groups exhibit some degree of stability in their behavior. However, successive samples of behavior from the same person are rarely identical in all pertinent respects. An individual's performances, products, and responses to sets of test questions vary in their quality or character from one occasion to another, even under strictly controlled conditions. This variation is reflected in the examinee's scores. The causes of this variability are generally unrelated to the purposes of measurement. An examinee may try harder, may make luckier guesses, be more alert, feel less anxious, or enjoy better health on one occasion than another. An examinee may have knowledge, experience, or understanding that is more relevant to some tasks than to others in the domain sampled by the test. Some individuals may exhibit less

variation in their scores than others, but no examinee is completely consistent. Because of this variation and, in some instances, because of subjectivity in the scoring process, an individual's obtained score and the average score of a group will always reflect at least a small amount of measurement error.

To say that a score includes a component of error implies that there is a hypothetical error-free value that characterizes an examinee at the time of testing. In classical test theory this error-free value is referred to as the person's *true score* for the test or measurement procedure. It is conceptualized as the hypothetical average score resulting from many repetitions of the test or alternate forms of the instrument. In statistical terms, the true score is a personal parameter and each observed score of an examinee is presumed to estimate this parameter. Under an approach to reliability estimation known as *generalizability theory*, a comparable concept is referred to as an examinee's *universe score*. Under *item response theory (IRT)*, a closely related concept is called an examinee's *ability or trait parameter*, though observed scores and trait parameters may be stated in different units. The hypothetical difference between an examinee's observed score on any particular measurement and the examinee's true or universe score for the procedure is called *measurement error*.

The definition of what constitutes a standardized test or measurement procedure has broadened significantly in recent years. At one time the cardinal features of most standardized tests were consistency of the test materials from examinee to examinee, close adherence to stipulated procedures for test administration, and use of prescribed scoring rules that could be applied with a high degree of consistency. These features were, in fact, what made a test "standardized," and they made meaningful norms possible. In employ-

ment settings and certification programs, flexible measurement procedures have been in use for many years. Individualized oral examinations, simulations, analyses of extended case reports, and performance in real-life settings such as clinics are now commonplace. In education, however, large-scale testing programs with a high degree of flexibility in test format and administrative procedures are a relatively recent development. In some programs cumulative portfolios of student work have been substituted for more traditional end-of-year tests of achievement. Other programs now allow examinees to choose their own topics to demonstrate their abilities. Still others permit or encourage small groups of examinees to work cooperatively in completing the test. A science examination, for example, might involve a team of high school students who conduct a study of the sources of pollution in local streams and prepare a report on their findings. Examinations of this kind raise complex issues regarding the domain represented by the test and about the generalizability of individual and group scores. Each step toward greater flexibility almost inevitably enlarges the scope and magnitude of measurement error. However, it is possible that some of the resultant sacrifices in reliability may reduce construct irrelevance or *construct underrepresentation in an assessment program*.

Characteristics and Implications of Measurement Error

Errors of measurement are generally viewed as random and unpredictable. They are conceptually distinguished from systematic errors, which may also affect performance of individuals or groups, but in a consistent rather than a *random manner*. For example, a systematic group error would occur as a result of differences in the difficulty of test forms that have not been adequately equated. When one test form is less difficult than another, examinees

who take the easier form may be expected to earn a higher average score than those who take the more difficult form. Such a difference would not be considered an error of measurement under most methods of quantifying and summarizing error, though generalizability theory would permit test form differences to be recognized as an error source.

The systematic factors that may differentially affect the performance of individual test takers are not as easily detected or overridden as those affecting groups. For example, some examinees experience levels of test anxiety that severely impair cognitive efficiency. The presence of such a condition can sometimes be recognized in an examinee, but the effect cannot be overcome by statistical adjustments. The individual systematic errors are not generally regarded as an element that contributes to unreliability. Rather, they constitute a source of construct-irrelevant variance and thus may detract from validity.

Important sources of measurement error may be broadly categorized as those rooted within the examinees and those external to them. Fluctuations in the level of an examinee's motivation, interest, or attention and the inconsistent application of skills are clearly internal factors that may lead to score inconsistencies. Differences among testing sites in their freedom from distractions, the random effects of scorer subjectivity, and variation in scorer standards are examples of external factors. The potency and importance of any particular source depend on the specific conditions under which the measures are taken, how performances are scored, and the interpretations made from the scores. A particular factor, such as the subjectivity in scoring, may be a significant source of measurement error in some assessments and a minor consideration in others.

Some changes in scores from one occasion to another, it should be noted, are not regarded as error, because they result, in part, from an intervention, learning, or maturation

that has occurred between the initial and final measures. The difference within an individual indicates, to some extent, the effects of the intervention or the extent of growth. In such settings, change per se constitutes the phenomenon of interest. The difference or the change score then becomes the measure to which reliability pertains.

Measurement error reduces the usefulness of measures. It limits the extent to which test results can be generalized beyond the particulars of a specific application of the measurement process. Therefore, it reduces the confidence that can be placed in any single measurement. Because random measurement errors are inconsistent and unpredictable, they cannot be removed from observed scores. However, their aggregate magnitude can be summarized in several ways, as discussed below.

Summarizing Reliability Data

Information about measurement error is essential to the proper evaluation and use of an instrument. This is true whether the measure is based on the responses to a specific set of questions, a portfolio of work samples, the performance of a task, or the creation of an original product. The ideal approach to the study of reliability entails independent replication of the entire measurement process. However, only a rough or partial approximation of such replication is possible in many testing situations, and investigation of measurement error may require special studies that depart from routine testing procedures. Nevertheless, it should be the goal of test developers to investigate test reliability as fully as practical considerations permit. No test developer is exempt from this responsibility.

The critical information on reliability includes the identification of the major sources of error, summary statistics bearing on the size of such errors, and the degree of generalizability of scores across alternate

forms, scorers, administrations, or other relevant dimensions. It also includes a description of the examinee population to whom the foregoing data apply, as the data may accurately reflect what is true of one population but misrepresent what is true of another. For example, a given reliability coefficient or estimated standard error derived from scores of a nationally representative sample may differ significantly from that obtained for a more homogeneous sample drawn from one gender, one ethnic group, or one community.

Reliability information may be reported in terms of variances or standard deviations of measurement errors, in terms of one or more coefficients, or in terms of IRT-based test information functions. The standard error of measurement is the standard deviation of a hypothetical distribution of measurement errors that arises when a given population is assessed via a particular test or procedure. The overall variance of measurement errors is actually a weighted average of the values that hold at various true score levels. The variance at a particular level is called a *conditional error variance* and its square root a *conditional standard error*. Traditionally, three broad categories of reliability coefficients have been recognized: (a) coefficients derived from the administration of parallel forms in independent testing sessions (alternate-form coefficients); (b) coefficients obtained by administration of the same instrument on separate occasions (test-retest or stability coefficients); and (c) coefficients based on the relationships among scores derived from individual items or subsets of the items within a test, all data accruing from a single administration (internal consistency coefficients). Where test scoring involves a high level of judgment, indexes of scorer consistency are commonly obtained. With the development of generalizability theory, the foregoing three categories may now be seen as special cases of a more general classification: generalizability coefficients.

Like traditional reliability coefficients, a *generalizability coefficient* is defined as the ratio of true or universe score variance to observed score variance. Unlike traditional approaches to the study of reliability, however, generalizability theory permits the researcher to specify and estimate the various components of true score variance, error variance, and observed score variance. Estimation is typically accomplished by the application of the techniques of analysis of variance. Of special interest are the separate numerical estimates of the components of overall error variance. Such estimates permit examination of the contribution of each source of error to the overall measurement process. The generalizability approach also makes possible the estimation of coefficients that apply to a wide variety of potential measurement designs.

The test information function, an important result of IRT, efficiently summarizes how well the test discriminates among individuals at various levels of the ability or trait being assessed. Under the IRT conceptualization, a mathematical function called the *item characteristic curve* or *item response function* is used as a model to represent the increasing proportion of correct responses to an item for groups at progressively higher levels of the ability or trait being measured. Given an adequate database, the parameters of the characteristic curve of each item in a test can be estimated. The test information function can then be approximated. This function may be viewed as a mathematical statement of the precision of measurement at each level of the given trait. Precision, in the IRT context, is analogous to the reciprocal of the conditional error variance of classical test theory.

Interpretation of Reliability Data

In general, reliability coefficients are most useful in comparing tests or measurement procedures, particularly those that yield scores in different units or metrics. However, such comparisons

are rarely straightforward. Allowance must be made for differences in the variability of the groups on which the coefficients are based, the techniques used to obtain the coefficients, the sources of error reflected in the coefficients, and the lengths of the instruments being compared in terms of testing time.

Generalizability coefficients and the many coefficients included under the traditional categories may appear to be interchangeable, but some convey quite different information from others. A coefficient in any given category may encompass errors of measurement from a highly restricted perspective, a very broad perspective, or some point between these extremes. For example, a coefficient may reflect error due to scorer inconsistencies but not reflect the variation that characterizes a succession of examinee performances or products. A coefficient may reflect only the internal consistency of item responses within an instrument and fail to reflect measurement error associated with day-to-day changes in examinee health, efficiency, or motivation.

It should not be inferred, however, that alternate-form or test-retest coefficients based on test administrations several days or weeks apart are always preferable to internal consistency coefficients. For many tests, internal consistency coefficients do not differ significantly from alternate-form coefficients. Where only one form of a test exists, retesting may result in an inflated correlation between the first and second scores due to idiosyncratic features of the test or to examinee recall of initial responses. Also, an individual's status on some attributes, such as mood or emotional state, may change significantly in a short period of time. In the assessment of such constructs the multiple measures that give rise to reliability estimates should be obtained within the short period in which the attribute remains stable. Therefore, for characteristics of this kind an internal consistency coefficient may be preferred.

The standard error of measurement is generally more relevant than the reliability coefficient once a measurement procedure has been adopted and interpretation of scores has become the user's primary concern. It should be noted that standard errors share some of the ambiguities which characterize reliability coefficients, and estimates may vary in their quality. Information about the precision of measurement at each of several widely spaced score levels—that is, conditional standard errors—is usually a valuable supplement to the single statistic for all score levels combined. Like reliability and generalizability coefficients, standard errors may reflect variation from many sources of error or only a few. For most purposes, a more comprehensive standard error is more informative than a less comprehensive value. However, there are many exceptions to this generalization. Practical constraints often preclude conduct of the kinds of studies that would yield estimates of the preferred standard errors.

Measurements derived from observations of behavior or evaluations of products are especially sensitive to a variety of error factors. These include evaluator biases and idiosyncrasies, scoring subjectivity, and intra-examinee factors that cause variation from one performance or product to another. The methods of generalizability theory are well suited to the investigation of the reliability of the scores on such measures. Estimates of the error variance associated with each specific source and with the interactions between sources indicate the extent to which examinee scores may be generalized to a population of scorers and to a universe of products or performances.

The interpretations of test scores may be broadly categorized as *relative* or *absolute*. Relative interpretations convey the standing of an individual or group within a reference population. Absolute interpretations relate the status of an individual or group to defined standards. These standards may originate in empirical data for one or more populations or

be based entirely on authoritative judgment. Different values of the standard error apply to the two types of interpretations.

The test information function can be perceived an alternative to traditional indices of measurement precision, but there are important distinctions that should be noted. Standard errors under classical test theory can be derived by several different approaches. These yield similar, but not identical, results. More significantly, standard errors, like reliability coefficients, may reflect a broad configuration of error factors or a restricted configuration, depending on the design of the reliability study. Test information functions, on the other hand, are limited to the restricted definition of measurement error that is associated with internal consistency reliabilities. In addition, under IRT several different mathematical models have been proposed and accepted as the basic form of the item characteristic curve. Adoption of one model rather than another can have a material effect on the derived test information function.

A final consideration has significant implications for both IRT and classical approaches to quantification of test score precision. It is this: Indices of precision depend on the scale in which they are reported. An index stated in terms of raw scores or the trait level estimates of IRT may convey a radically different perception of reliability than the same index restated in terms of derived scores. This same contrast may hold for conditional standard errors. In terms of the basic score scale, precision may appear to be high at one score level, low at another. But when the conditional standard errors are restated in units of derived scores, such as grade equivalents or standard scores, quite different trends in comparative precision may emerge. Therefore, measurement precision under both theories very strongly depends on the scale in which test scores are reported and interpreted.

Precision and consistency in measurement are always desirable. However, the need

for precision increases as the consequences of decisions and interpretations grow in importance. If a decision can and will be corroborated by information from other sources or if an erroneous initial decision can be quickly corrected, scores with modest reliability may suffice. But if a test score leads to a decision that is not easily reversed, such as rejection or admission of a candidate to a professional school or the decision by a jury that a serious injury was sustained, the need for a high degree of precision is much greater.

Where the purpose of measurement is classification, some measurement errors are more serious than others. An individual who is far above or far below the value established for pass/fail or for eligibility for a special program can be mismeasured without serious consequences. Mismeasurement of examinees whose true scores are close to the cut score is a more serious concern. The techniques used to quantify reliability should recognize these circumstances. This can be done by reporting the conditional standard error in the vicinity of the critical value.

Some authorities have proposed that a semantic distinction be made between "reliability of scores" and "degree of agreement in classification." The former term would be reserved for analysis of score variation under repeated measurement. The term *classification consistency* or *inter-rater agreement*, rather than *reliability*, would be used in discussions of consistency of classification. Adoption of such usage would make it clear that the importance of an error of any given size depends on the proximity of the examinee's score to the cut score. However, it should be recognized that the degree of consistency or agreement in examinee classification is specific to the cut score employed and its location within the score distribution.

Average scores of groups, when interpreted as measures of program effectiveness, involve error factors that are not identical to those that operate at the individual level. For

large groups, the positive and negative measurement errors of individuals may average out almost completely in group means. However, the sampling errors associated with the random sampling of persons who are tested for purposes of program evaluation are still present. This component of the variation in the mean achievement of school classes from year to year or in the average expressed satisfaction of successive samples of the clients of a program may constitute a potent source of error in program evaluations. It can be a significant source of error in inferences about programs even if there is a high degree of precision in individual test scores. Therefore, when an instrument is used to make group judgments, reliability data must bear directly on the interpretations specific to groups. Standard errors appropriate to individual scores are not appropriate measures of the precision of group averages. A more appropriate statistic is the standard error of the observed score means. Generalizability theory can provide more refined indices when the sources of measurement error are numerous and complex.

Typically, developers and distributors of tests have primary responsibility for obtaining and reporting evidence of reliability or test information functions. The user must have such data to make an informed choice among alternative measurement approaches and will generally be unable to conduct reliability studies prior to operational use of an instrument. In some instances, however, local users of a test or procedure must accept at least partial responsibility for documenting the precision of measurement. This obligation holds when one of the primary purposes of measurement is to rank or classify examinees within the local population. It also holds when users must rely on local scorers who are trained to use the scoring rubrics provided by the test developer. In such settings, local factors may materially affect the magnitude of error variance and observed score variance. Therefore, the reliability of

scores may differ appreciably from that reported by the developer.

The reporting of reliability coefficients alone, with little detail regarding the methods used to estimate the coefficient, the nature of the group from which the data were derived, and the conditions under which the data were obtained constitutes inadequate documentation. General statements to the effect that a test is "reliable" or that it is "sufficiently reliable to permit interpretations of individual scores" are rarely, if ever, acceptable. It is the user who must take responsibility for determining whether or not scores are sufficiently trustworthy to justify anticipated uses and interpretations. Of course, test constructors and publishers are obligated to provide sufficient data to make informed judgments possible.

As the foregoing comments emphasize, there is no single, preferred approach to quantification of reliability. No single index adequately conveys all of the relevant facts. No one method of investigation is optimal in all situations, nor is the test developer limited to a single approach for any instrument. The choice of estimation techniques and the minimum acceptable level for any index remain a matter of professional judgment.

Although reliability is discussed here as an independent characteristic of test scores, it should be recognized that the level of reliability of scores has implications for the validity of score interpretations. Reliability data ultimately bear on the repeatability of the behavior elicited by the test and the consistency of the resultant scores. The data also bear on the consistency of classifications of individuals derived from the scores. To the extent that scores reflect random errors of measurement, their potential for accurate prediction of criteria, for beneficial examinee diagnosis, and for wise decision making is limited. Relatively unreliable scores, in conjunction with other convergent information, may sometimes be of value to a test user, but the level of a score's reliability places limits on its unique contribution to validity for all purposes.

Standard 2.1

For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported.

Comment: It is not sufficient to report estimates of reliabilities and standard errors of measurement only for total scores when subscores are also interpreted. The form-to-form and day-to-day consistency of total scores on a test may be acceptably high, yet subscores may have unacceptably low reliability. For all scores to be interpreted, users should be supplied with reliability data in enough detail to judge whether scores are precise enough for the users' intended interpretations. Composites formed from selected subtests within a test battery are frequently proposed for predictive and diagnostic purposes. Users need information about the reliability of such composites.

Standard 2.2

The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original scale units and in units of each derived score recommended for use in test interpretation.

Comment: The most common derived scores include standard scores, grade or age equivalents, and percentile ranks. Because raw scores on norm-referenced tests are only rarely interpreted directly, standard errors in derived score units are more helpful to the typical test user. A confidence interval for an examinee's true score, universe score, or percentile rank serves much the same purpose as a standard error and can be used as an alternative approach to convey reliability information. The implications of the standard error of measurement are especially important in situations where decisions cannot be postponed and corroborative sources of information are limited.

STANDARDS

Standard 2.3

When test interpretation emphasizes differences between two observed scores of an individual or two averages of a group, reliability data, including standard errors, should be provided for such differences.

Comment: Observed score differences are used for a variety of purposes. Achievement gains are frequently the subject of inferences for groups as well as individuals. Differences between verbal and performance scores of intelligence and scholastic ability tests are often employed in the diagnosis of cognitive impairment and learning problems. Psycho-diagnostic inferences are frequently drawn from the differences between subtest scores. Aptitude and achievement batteries, interest inventories, and personality assessments are commonly used to identify and quantify the relative strengths and weaknesses or the pattern of trait levels of an examinee. When the interpretation of test scores centers on the peaks and valleys in the examinee's test score profile, the reliability of score differences for all pairs of scores is critical.

Standard 2.4

Each method of quantifying the precision or consistency of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select examinees for reliability analyses and descriptive statistics on these samples should be reported.

Comment: Information on the method of subject selection, sample sizes, means, standard deviations, and demographic characteristics of the groups helps users judge the extent to which reported data apply to their own examinee populations. If the test-retest or alternate-form approach is used, the interval between testings should be indicated. Because there are many ways of estimating reliability,

each influenced by different sources of measurement error, it is unacceptable to say simply, "The reliability of test X is .90." A better statement would be, "The reliability coefficient of .90 reported for scores on test X was obtained by correlating scores from forms A and B administered on successive days. The data were based on a sample of 400 10th-grade students from five middle-class suburban schools in New York State. The demographic breakdown of this group was as follows:"

Standard 2.5

A reliability coefficient or standard error of measurement based on one approach should not be interpreted as interchangeable with another derived by a different technique unless their implicit definitions of measurement error are equivalent.

Comment: Internal consistency, alternate-form, test-retest, and generalizability coefficients should not be considered equivalent, as each may incorporate a unique definition of measurement error. Error variances derived via item response theory may not be equivalent to error variances estimated via other approaches. Test developers should indicate the sources of error that are reflected in or ignored by the reported reliability indices.

Standard 2.6

If reliability coefficients are adjusted for restriction of range or variability, the adjustment procedure and both the adjusted and unadjusted coefficients should be reported. The standard deviations of the group actually tested and of the target population, as well as the rationale for the adjustment, should be presented.

Comment: Application of a correction for restriction in variability presumes that the available sample is not representative of the test-taker population to which users might be expected to generalize. The rationale for the

correction should consider the appropriateness of such a generalization. Adjustment formulas that presume constancy in the standard error across score levels should not be used unless constancy can be defended.

Standard 2.7

When subsets of items within a test are dictated by the test specifications and can be presumed to measure partially independent traits or abilities, reliability estimation procedures should recognize the multifactor character of the instrument.

Comment: The total score on a test that is clearly multifactor in nature should be treated as a composite score. If an internal consistency estimate of total score reliability is obtained by the split-halves procedure, the halves should be parallel in content and statistical characteristics. Stratified coefficient alpha should be used rather than the more familiar nonstratified coefficient.

Standard 2.8

Test users should be informed about the degree to which rate of work may affect examinee performance.

Comment: It is not possible to state, in general, whether reliability coefficients will increase or decrease when rate of work becomes an important source of systematic variance. Rate of work, as an examinee trait, may be more stable or less stable from occasion to occasion than the other factors the test is designed to measure. Because speededness has differential effects on various estimates, information on speededness is helpful in interpreting reported coefficients.

The importance of the speed factor can sometimes be inferred from analyses of item responses and from observations by examiners during test administrations conducted for reliability analyses. The distribution of “last item attempted” and increases in the frequen-

cy of omitted responses toward the end of a test are also highly informative, though not conclusive, evidence regarding speededness. A decline in the proportion of correct responses, beyond that attributable to increasing item difficulty, may indicate that some examinees were responding randomly. With computer-administered tests, abnormally fast item response times, particularly toward the end of the test, may also suggest that examinees were responding randomly. In the case of constructed-response exercises, including essay questions, the completeness of the responses may suggest that time constraints had little effect on early items but a significant effect on later items. Introduction of a speed factor into what might otherwise be a power test may have a marked effect on alternate-form and test-retest reliabilities. A shift from a paper-and-pencil format to a computer-administered format may affect test speededness.

Standard 2.9

When a test is designed to reflect rate of work, reliability should be estimated by the alternate-form or test-retest approach, using separately timed administrations.

Comment: Split-half coefficients based on separate scores from the odd-numbered and even-numbered items are known to yield inflated estimates of reliability for highly speeded tests. Coefficient alpha and other internal consistency coefficients may also be biased, though the size of the bias is not as clear as that for the split-halves coefficient.

Standard 2.10

When subjective judgment enters into test scoring, evidence should be provided on both inter-rater consistency in scoring and within-examinee consistency over repeated measurements. A clear distinction should be made among reliability data based on (a) independent panels of raters scoring the same perform-

STANDARDS

RELIABILITY AND ERRORS OF MEASUREMENT / PART I

ances or products, (b) a single panel scoring successive performances or new products, and (c) independent panels scoring successive performances or new products.

Comment: Task-to-task variations in the quality of an examinee's performance and rater-to-rater inconsistencies in scoring represent independent sources of measurement error. Reports of reliability studies should make clear which of these sources are reflected in the data. Where feasible, the error variances arising from each source should be estimated. Generalizability studies and variance component analyses are especially helpful in this regard. These analyses can provide separate error variance estimates for tasks within examinees, for judges, and for occasions within the time period of trait stability. Information should be provided on the qualifications of the judges used in reliability studies.

Inter-rater or inter-observer agreement may be particularly important for ratings and observational data that involve subtle discriminations. It should be noted, however, that when raters evaluate positively correlated characteristics, a favorable or unfavorable assessment of one trait may color their opinions of other traits. Moreover, high inter-rater consistency does not imply high examinee consistency from task to task. Therefore, internal consistency within raters and inter-rater agreement do not guarantee high reliability of examinee scores.

Standard 2.11

If there are generally accepted theoretical or empirical reasons for expecting that reliability coefficients, standard errors of measurement, or test information functions will differ substantially for various subpopulations, publishers should provide reliability data as soon as feasible for each major population for which the test is recommended.

Comment: If test score interpretation involves inferences within subpopulations as well as within the general population, reliability data should be provided for both the subpopulations and the general population. Test users who work exclusively with a specific cultural group or with individuals who have a particular disability would benefit from an estimate of the standard error for such a subpopulation. Some groups of test takers—pre-school children, for example—tend to respond to test stimuli in a less consistent fashion than do older children.

Standard 2.12

If a test is proposed for use in several grades or over a range of chronological age groups and if separate norms are provided for each grade or each age group, reliability data should be provided for each age or grade population, not solely for all grades or ages combined.

Comment: A reliability coefficient based on a sample of examinees spanning several grades or a broad range of ages in which average scores are steadily increasing will generally give a spuriously inflated impression of reliability. When a test is intended to discriminate within age or grade populations, reliability coefficients and standard errors should be reported separately for each population.

Standard 2.13

If local scorers are employed to apply general scoring rules and principles specified by the test developer, local reliability data should be gathered and reported by local authorities when adequate size samples are available.

Comment: For example, many statewide testing programs depend on local scoring of essays, constructed-response exercises, and performance tests. Reliability analyses bear on the possibility that additional training of scorers is needed and, hence, should be an integral part of program monitoring.

Standard 2.14

Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.

Comment: Estimation of conditional standard errors is usually feasible even with the sample sizes that are typically used for reliability analyses. If it is assumed that the standard error is constant over a broad range of score levels, the rationale for this assumption should be presented.

Standard 2.15

When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure, using the same form or alternate forms of the instrument.

Comment: When a test or composite is used to make categorical decisions, such as pass/fail, the standard error of measurement at or near the cut score has important implications for the trustworthiness of these decisions. However, the standard error cannot be translated into the expected percentage of consistent decisions unless assumptions are made about the form of the distributions of measurement errors and true scores. It is preferable that this percentage be estimated directly through the use of a repeated-measurements approach if consistent with the requirements of test security and if adequate samples are available.

Standard 2.16

In some testing situations, the items vary from examinee to examinee—through random selection from an extensive item pool or application

of algorithms based on the examinee's level of performance on previous items or preferences with respect to item difficulty. In this type of testing, the preferred approach to reliability estimation is one based on successive administrations of the test under conditions similar to those prevailing in operational test use.

Comment: Varying the set of items presented to each examinee is an acceptable procedure in some settings. If this approach is used, reliability data should be appropriate to this procedure. Estimates of standard errors of ability scores can be computed through the use of IRT and reported routinely as part of the adaptive testing procedure. However, those estimates are not an adequate substitute for estimates based on successive administrations of the adaptive test, nor do they bear on the issue of stability over short intervals. IRT estimates are contingent on the adequacy of both the item parameter estimates and the item response models adopted in the theory. Estimates of reliabilities and standard errors of measurement based on the administration and analysis of alternate forms of an adaptive test reflect errors associated with the entire measurement process. The alternate-form estimates provide an independent check on the magnitude of the errors of measurement specific to the adaptive feature of the testing procedure.

Standard 2.17

When a test is available in both long and short versions, reliability data should be reported for scores on each version, preferably based on an independent administration of each.

Comment: Some tests and test batteries are published in both a "full-length" version and a "survey" or "short" version. In many applications the Spearman-Brown formula will satisfactorily approximate the reliability of one of these from data based on the other. However, context effects are commonplace in tests of

STANDARDS

RELIABILITY AND ERRORS OF MEASUREMENT / PART I

maximum performance. Also, the short version of a standardized test often comprises a nonrandom sample of items from the full-length version. Therefore, the shorter version may be more reliable or less reliable than the Spearman-Brown projections from the full-length version. The reliability of scores on each version is best evaluated through an independent administration of each, using the designated time limits.

Standard 2.18

When significant variations are permitted in test administration procedures, separate reliability analyses should be provided for scores produced under each major variation if adequate sample sizes are available.

Comment: To accommodate examinees with disabilities, test publishers might authorize modifications in the procedures and time limits that are specified for the administration of the paper-and-pencil edition of a test. In some cases, modified editions of the test itself may be provided. For example, tape-recorded versions for use in a group setting or with individual equipment may be used to test examinees who exhibit reading disabilities or attention deficits. If such modifications can be employed with test takers who are not disabled, insights can be gained regarding the possible effects on test scores of these non-standard administrations.

Standard 2.19

When average test scores for groups are used in program evaluations, the groups tested should generally be regarded as a sample from a larger population, even if all examinees available at the time of measurement are tested. In such cases the standard error of the group mean should be reported, as it reflects variability due to sampling of examinees as well as variability due to measurement error.

Comment: The graduating seniors of a liberal arts college, the current clients of a social service agency, and analogous groups exposed to a program of interest typically constitute a sample in a longitudinal sense. Presumably, comparable groups from the same population will recur in future years, given static conditions. The factors leading to uncertainty in conclusions about program effectiveness arise from the sampling of persons as well as measurement error. Therefore, the standard error of the mean observed score, reflecting variation in both true scores and measurement errors, represents a more realistic standard error in this setting. Even this value may underestimate the variability of group means over time. In many settings, the static conditions assumed under random sampling of persons do not prevail.

Standard 2.20

When the purpose of testing is to measure the performance of groups rather than individuals, a procedure frequently used is to assign a small subset of items to each of many subsamples of examinees. Data are aggregated across subsamples and item subsets to obtain a measure of group performance. When such procedures are used for program evaluation or population descriptions, reliability analyses must take the sampling scheme into account.

Comment: This type of measurement program is termed *matrix sampling*. It is designed to reduce the time demanded of individual examinees and to increase the total number of items on which data are obtained. This testing approach provides the same type of information about group performances that would accrue if all examinees could respond to all exercises in the item pool. Reliability statistics must be appropriate to the sampling plan used with respect to examinees and items.

3. TEST DEVELOPMENT AND REVISION

Background

Test development is the process of producing a measure of some aspect of an individual's knowledge, skill, ability, interests, attitudes, or other characteristics by developing items and combining them to form a test, according to a specified plan. Test development is guided by the stated purpose(s) of the test and the intended inferences to be made from the test scores. The test development process involves consideration of content, format, the context in which the test will be used, and the *potential consequences of using the test*. Test development also includes specifying conditions for administering the test, determining procedures for scoring the test performance, and reporting the scores to test takers and test users. This chapter focuses primarily on the following aspects of test development: stating the purpose(s) of the test, defining a framework for the test, developing test specifications, developing and evaluating items and their associated scoring procedures, assembling the test, and revising the test. The first section describes the test development process that begins with a statement of the purpose(s) of the test and culminates with the assembly of the test. The second section addresses several special considerations in test development, including considerations in delineating the test framework and in developing performance assessments. The chapter concludes with a discussion on test revision. Issues bearing on validity, reliability, and fairness are interwoven within the stages of test development. Each of these topics is addressed comprehensively in other chapters of the *Standards*: validity in chapter 1, reliability in chapter 2, and aspects of fairness in chapters 7, 8, 9, and 10. Additional material on test administration and scoring, and on reporting scores and results, is provided in chapter 5. Chapter 4 discusses score scales, and the focus of chapter 6 is test documents.

Test Development

The process of developing educational and psychological tests commonly begins with a statement of the purpose(s) of the test and the construct or content domain to be measured. Tests of the same construct or domain can differ in important ways, because a number of decisions must be made as the test is developed. It is helpful to consider the four phases leading from the original statement of purpose(s) to the final product: (a) delineation of the purpose(s) of the test and the scope of the construct or the extent of the domain to be measured; (b) development and evaluation of the test specifications; (c) development, field testing, evaluation, and selection of the items and scoring guides and procedures; and (d) assembly and evaluation of the test for operational use. What follows is a description of typical test development procedures, though there may be sound reasons that some of these steps are followed in some settings and not in others.

The first step is to extend the original statement of purpose(s), and the construct or content domain being considered, into a framework for the test that describes the extent of the domain, or the scope of the construct to be measured. The test framework, therefore, delineates the aspects (e.g., content, skills, processes, and diagnostic features) of the construct or domain to be measured. For example, "Does eighth-grade mathematics include algebra?" "Does verbal ability include text comprehension as well as vocabulary?" "Does self-esteem include both feelings and acts?" The delineation of the test framework can be guided by theory or an analysis of the content domain or job requirements as in the case of many licensing and employment tests. The test framework serves as a guide to subsequent test evaluation. The chapter on validity provides a more thorough discussion of the relationships among the construct or content domain, the test framework, and the purpose(s) of the test.

Once decisions have been made about what the test is to measure, and what its scores are intended to convey, the next step is to design the test by establishing test specifications. The test specifications delineate the format of items, tasks, or questions; the response format or conditions for responding; and the type of scoring procedures. The specifications may indicate the desired psychometric properties of items, such as difficulty and discrimination, as well as the desired test properties such as test difficulty, inter-item correlations, and reliability. The test specifications may also include such factors as time restrictions, characteristics of the intended population of test takers, and procedures for administration. All subsequent test development activities are guided by the test specifications.

Test specifications will include, at least implicitly, an indication of whether the test scores will be primarily norm-referenced or criterion-referenced. When scores are norm-referenced, relative score interpretations are of primary interest. A score for an individual or for a definable group is ranked within one or more distributions of scores or compared to the average performance of test takers for various reference populations (e.g., based on age, grade, diagnostic category, or job classification). When scores are criterion-referenced, absolute score interpretations are of primary interest. The meaning of such scores does not depend on rank information. Rather, the test score conveys directly a level of competence in some defined criterion domain. Both relative and absolute interpretations are often used with a given test, but the test developer determines which approach is most relevant for that test.

The nature of the item and response formats that may be specified depends on the purposes of the test and the defined domain of the test. Selected-response formats, such as multiple-choice items, are suitable for many purposes of testing. The test specifications indicate how many alternatives are to be used

for each item. Other purposes may be more effectively served by a short constructed-response format. Short-answer items require a response of no more than a few words. Extended-response formats require the test taker to write a more extensive response of one or more sentences or paragraphs. Performance assessments often seek to emulate the context or conditions in which the intended knowledge or skills are actually applied. One type of performance assessment, for example, is the standardized job or work sample. A task is presented to the test taker in a standardized format under standardized conditions. Job or work samples might include, for example, the assessment of a practitioner's ability to make an accurate diagnosis and recommend treatment for a defined condition, a manager's ability to articulate goals for an organization, or a student's proficiency in performing a science laboratory experiment.

All types of items require some indication of how to score the responses. For selected-response items, one alternative is considered the correct response in some testing programs. In other testing programs, the alternatives may be weighted differentially. For short-answer items, a list of acceptable alternatives may suffice; extended-response items need more detailed rules for scoring, sometimes called *scoring rubrics*. Scoring rubrics specify the criteria for evaluating performance and may vary in the degree of judgment entailed, in the number of score levels, and in other ways. It is common practice for test developers to provide scorers with examples of performances at each of the score levels to help clarify the criteria.

For extended-response items, including performance tasks, two major types of scoring procedures are used: analytic and holistic. Both of the procedures require explicit performance criteria that reflect the test framework. However, the approaches differ in the degree of detail provided in the evaluation report. Under the analytic scoring procedure, each critical dimension of the performance criteria is judged independently, and separate scores are obtained

PART I / TEST DEVELOPMENT AND REVISION

for each of these dimensions in addition to an overall score. Under the holistic scoring procedure, the same performance criteria may implicitly be considered, but only one overall score is provided. Because the analytic procedure provides information on a number of critical dimensions, it potentially provides valuable information for diagnostic purposes and lends itself to evaluating strengths and weaknesses of test takers. In contrast, the holistic procedure may be preferable when an overall judgment is desired and when the skills being assessed are complex and highly interrelated. Regardless of the type of scoring procedure, designing the items and developing the scoring rubrics and procedures is an integrated process.

A participatory approach may be used in the design of items, scoring rubrics, and sometimes the scoring process itself. Many interested persons (e.g., practitioners, teachers) may be involved in developing items and scoring rubrics, and/or evaluating the subsequent performances. If a participatory approach is used, participants' knowledge about the domain being assessed and their ability to apply the scoring rubrics are of critical importance. Equally important, for those involved in developing tests and evaluating performances, is their familiarity with the nature of the population being tested. Relevant characteristics of the population being tested may include the typical range of expected skill levels, their familiarity with the response modes required of them, and the primary language they use.

The test developer usually assembles an item pool that consists of a larger set of items than what is required by the test specifications. This allows for the test developer to select a set of items for the test that meet the test specifications. The quality of the items is usually ascertained through item review procedures and pilot testing. Items are reviewed for content quality, clarity and lack of ambiguity. Items sometimes are reviewed for sensitivity to gender or cultural issues. An attempt is generally made to avoid words and topics

that may offend or otherwise disturb some test takers, if less offensive material is equally useful. Often, a field test is developed and administered to a group of test takers who are somewhat representative of the target population for the test. The field test helps determine some of the psychometric properties of the test items, such as an item's difficulty and ability to discriminate among test takers of different standing on the scale. Ongoing testing programs often pretest items by inserting them into existing tests. Those items are not used in obtaining test scores of the test takers, but the item responses provide useful data for test development.

The next step in test development is to assemble items into a test or to identify an item pool for an adaptive test. The test developer is responsible for ensuring that the items selected for the test meet the requirements of the test specifications. Depending upon the purpose(s) of the test, relevant considerations in item selection may include the content quality and scope, the weighting of items and subdomains, and the appropriateness of the items selected for the intended population of test takers. Often test developers will specify the distribution of psychometric indices of the items to be included in the test. For example, the specified distribution of item difficulty indices for a selection test would differ from the distribution specified for a general achievement test. When psychometric indices of the items are estimated using item response theory (IRT), the fit of the model to the data is also evaluated. This is accomplished by evaluating the extent to which the assumptions underlying the item response model (e.g., unidimensionality, local independence, speededness, and equality of slope parameters) are satisfied.

The test developer is also responsible for ensuring that the scoring procedures are consistent with the purpose(s) of the test and facilitate meaningful score interpretation. The nature of the intended score interpretations

will determine the importance of psychometric characteristics of items in the test construction process. For example, indices of item difficulty and discrimination, and inter-item correlations, may be particularly important when relative score interpretations are intended. In the case of relative score interpretations, good discrimination among test takers at all points along the construct continuum is desirable. It is important, however, that the test specifications are not compromised when optimizing the distribution of these indices. In the case of absolute score interpretations, different criteria apply. In this case, the extent to which the relevant domain has been adequately represented is important even if many of the items are relatively easy or nondiscriminating within a relevant population. It is important, however, to assure the quality of the content of relatively easy or nondiscriminating items. If cut scores are necessary for score interpretation in criterion-referenced programs, the level of item discrimination constitutes critical information primarily in the vicinity of the cut scores. Because of these differences in test development procedures, tests designed to facilitate one type of interpretation function less effectively for other types of interpretation. Given appropriate test design and supporting evidence, however, scores arising from some norm-referenced programs may provide reasonable absolute score interpretations and scores arising from some criterion-referenced programs may provide reasonable relative score interpretations.

When evaluating the quality of the items in the item pool and the test itself, test developers often conduct studies of differential item functioning (see chapter 7). Differential item functioning is said to exist when test takers of approximately equal ability on the targeted construct or content domain differ in their responses to an item according to their group membership. In theory, the ultimate goal of such studies is to identify construct-irrelevant aspects of item content, item format,

or scoring criteria that may differentially affect test scores of one or more groups of test takers. When differential item functioning is detected, test developers try to identify plausible explanations for the differences, and then they may replace or revise items that give rise to group differences if construct irrelevance is deemed likely. However, at this time, there has been little progress in discerning the cause or substantive themes that account for differential item functioning on a group basis. Items for which the differential item functioning index is significant may constitute valid measures of an element of the intended domain and differ in no way from other items that show nonsignificant indexes. When the differential item functioning index is significant, the test developer must take care that any replacement items or item revisions do not compromise the test specifications.

When multiple forms of a test are prepared, the test specifications govern each of the forms. Also, when an item pool is developed for a computerized adaptive test, the specifications refer both to the item pool and to the rules or procedures by which the individual item sets are created for each test taker. Some of the attractive features of computerized adaptive tests, such as tailoring the difficulty level of the items to the test taker's ability, place additional constraints on the design of such tests. In general, a large number of items is needed for a computerized adaptive test to ensure that each tailored item set meets the requirements of the test specifications. Further, tests often are developed in the context of larger systems or programs. Multiple item sets, for example, may be created for use with different groups of test takers or on different testing dates. Last, when a short form of a test is prepared, the test specifications of the original test govern the short form. Differences in the test specifications and the psychometric properties of the short form and the original test will affect the interpretation of the scores derived from the short

form. In any of these cases, the same fundamental methods and principles of test development apply.

Special Considerations in Test Development

This section elaborates on several topics discussed above. First, considerations in delineating the framework for the test are discussed. Following this, considerations in the development of performance assessments and portfolios are addressed.

Delineating the Framework for the Test

The scenario presented above outlines what is often done to develop a test. However, the activities do not always happen in a rigid sequence. There is often a subtle interplay between the process of conceptualizing a construct or content domain and the development of a test of that construct or domain. The framework for the test provides a description of how the construct or domain will be represented. The procedures used to develop items and scoring rubrics and to examine item characteristics may often contribute to clarifying the framework. The extent to which the framework is defined a priori is dependent on the testing application. In many testing applications, a well-defined framework and detailed test specifications guide the development of items and their associated scoring rubrics and procedures. In some areas of psychological measurement, test development may be less dependent on an a priori defined framework and may rely more on a data-based approach that results in an empirically derived definition of the framework. In such instances, items are selected primarily on the basis of their empirical relationship with an external criterion, their relationships with one another, or their power to discriminate among groups of individuals. For example, construction of a selection test for sales personnel might be guided by the corre-

lations of item scores with productivity measures of current sales personnel or a measure of client satisfaction might be assembled from those items in an item pool that correlate most highly with customer loyalty. Similarly, an inventory to help identify different patterns of psychopathology might be developed using patients from different diagnostic subgroups. When test development relies on a data-based approach, it is likely that some items will be selected based on chance occurrences in the data. Cross-validation studies are routinely conducted to determine the tendency to select items by chance, which involves administering the test to a comparable sample.

In many testing applications, the framework for the test is specified initially and this specification subsequently guides the development of items and scoring procedures. Empirical relationships may then be used to inform decisions about retaining, rejecting, or modifying items. Interpretations of scores from tests developed by this process have the advantage of a logical/theoretical and an empirical foundation for the underlying dimensions represented by the test.

PERFORMANCE ASSESSMENTS

One distinction between performance assessments and other forms of tests has to do with the type of response that is required from the test takers. Performance assessments require the test takers to carry out a process such as playing a musical instrument or tuning a car's engine or to produce a product such as a written essay. Performance assessments generally require the test takers to demonstrate their abilities or skills in settings that closely resemble real-life settings. For example, an assessment of a psychologist in training may require the test taker to interview a client, choose appropriate tests, and arrive at diagnosis and *plan for therapy*. Performance assessments are diverse in nature and can be product-based as well as behavior-based. Because performance assessments typically consist of a small num-

ber of tasks, establishing the extent to which the results can be generalized to the broader domain is particularly important. The use of test specifications will contribute to tasks being developed so as to systematically represent the critical dimensions to be assessed, leading to a more comprehensive coverage of the domain than what would occur if test specifications were not used. Further, both logical and empirical evidence are important to document the extent to which performance assessments—tasks as well as scoring criteria—reflect the processes or skills that are specified by the domain definition. When tasks are designed to elicit complex cognitive processes, logical analyses of the tasks and both logical and empirical analyses of the test takers' performances on the tasks provide necessary validity evidence.

PORTFOLIOS

A unique type of performance assessment is an individual portfolio. Portfolios are systematic collections of work or educational products typically collected over time. Like other assessment procedures, the design of portfolios is dependent on the purpose. Typical purposes include judgment of the improvement in job or educational performance and evaluation of the eligibility for employment, promotion, or graduation. A well-designed portfolio specifies the nature of the work that is to be put into the portfolio. The portfolio may include entries such as representative products, the best work of the test taker, or indicators of progress. For example, in an employment setting involving promotion, employees may be instructed to include their best work or products. Alternatively, if the purpose is to judge a student's educational growth, students may be asked to provide evidence of improvement with respect to particular competencies or skills. They may also be requested to provide justifications for the choices. Still other methods may include the use of videotapes, exhibitions, demonstrations, simulations, and so on.

In employment settings, employees may be involved in the selection of their work and prod-

ucts that demonstrate their competencies for promotion purposes. Analogously, in educational applications, students may participate in the selection of some of their work and the products to be included in their portfolios as well as in the evaluation of the materials. The specifications for the portfolio indicate who is responsible for selecting its contents. For example, the specifications may state that the test taker, the examiner, or both parties working together should be involved in the selection of the contents of the portfolio. The particular responsibilities of each party are delineated in the specifications. The more standardized the contents and procedures of administration, the easier it is to establish comparability of portfolio-based scores. Regardless of the methods used, all performance assessments are evaluated by the same standards of technical quality as other forms of tests.

Test Revisions

Tests and their supporting documents (e.g., test manuals, technical manuals, user's guides) are reviewed periodically to determine whether revisions are needed. Revisions or amendments are necessary when new research data, significant changes in the domain, or new conditions of test use and interpretation would either improve the validity of interpretations of the test scores or suggest that the test is no longer fully appropriate for its intended use. As an example, tests are revised if the test content or language has become outdated and, therefore, may subsequently affect the validity of the test score interpretations. Revisions to test content are also made to ensure the confidentiality of the test. It should be noted, however, that outdated norms may not have the same implications for revisions as an outdated test. For example, it may be necessary to update the norms for an achievement test after a period of rising or falling achievement in the norming population, or when there are changes in the test-taking population, but the test content itself may continue to be as relevant as it was when the test was developed.

Standard 3.1

Tests and testing programs should be developed on a sound scientific basis. Test developers and publishers should compile and document adequate evidence bearing on test development.

Standard 3.2

The purpose(s) of the test, definition of the domain, and the test specifications should be stated clearly so that judgments can be made about the appropriateness of the defined domain for the stated purpose(s) of the test and about the relation of items to the dimensions of the domain they are intended to represent.

Comment: The adequacy and usefulness of test interpretations depend on the rigor with which the purposes of the test and the domain represented by the test have been defined and explicated. The domain definition should be sufficiently detailed and delimited to show clearly what dimensions of knowledge, skill, processes, attitude, values, emotions, or behavior are included and what dimensions are excluded. A clear description will enhance accurate judgments by reviewers and others about the congruence of the defined domain and the test items.

Standard 3.3

The test specifications should be documented, along with their rationale and the process by which they were developed. The test specifications should define the content of the test, the proposed number of items, the item formats, the desired psychometric properties of the items, and the item and section arrangement. They should also specify the amount of time for testing, directions to the test takers, procedures to be used for test administration and scoring, and other relevant information.

Comment: Professional judgment plays a major role in developing the test specifications. The specific procedures used for developing the specifications depend on the purposes of the test. For example, in developing licensure and certification tests, practice analyses or job analyses usually provide the basis for defining the test specifications, and job analyses primarily serve this function for employment tests. For achievement tests to be given at the end of a course, the test specifications should be based on an outline of course content and goals. Whereas, for placement tests, it may be necessary to examine the required entry knowledge and skills for several courses.

Standard 3.4

The procedures used to interpret test scores, and, when appropriate, the normative or standardization samples or the criterion used should be documented.

Comment: Test specifications may indicate that the intended score interpretations are for absolute or relative score interpretations, or both. In relative score interpretations the status of an individual (or group) is determined by comparing the score (or mean score) to the performance of others in one or more defined populations. In absolute score interpretations, the score or average is assumed to reflect directly a level of competence or mastery in some defined criterion domain. Tests designed to facilitate one type of interpretation function less effectively for other types of interpretations. Given appropriate test design and adequate supporting data, however, scores arising from norm-referenced testing programs may provide reasonable absolute score interpretations and scores arising from criterion-referenced programs may provide reasonable relative score interpretations.

Standard 3.5

When appropriate, relevant experts external to the testing program should review the test specifications. The purpose of the review, the

STANDARDS

TEST DEVELOPMENT AND REVISION / PART I

process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.

Comment: Expert review of the test specifications may serve many useful purposes such as helping to assure content quality and representativeness. The expert judges may include individuals representing defined populations of concern to the test specifications. For example, if the test is related to ethnic minority concerns, the expert review typically includes members of appropriate ethnic minority groups or experts on minority group issues.

Standard 3.6

The type of items, the response formats, scoring procedures, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that intended inferences from test scores are equally valid for members of different groups of test takers. The test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.

Comment: Expert judges may be asked to identify material likely to be inappropriate, confusing, or offensive for groups in the test-taking population. For example, judges may be asked to identify whether lack of exposure to problem contexts in mathematics word problems may be of concern for some groups of students. Various groups of test takers can be defined by characteristics such as age, ethnicity, culture, gender, disability, or demographic region. There is limited evidence, however, that expert reviews alleviate problems with bias in testing (see chapter 7).

Standard 3.7

The procedures used to develop, review, and try out items, and to select items from the item pool should be documented. If the items were classified into different categories or subtests according to the test specifications, the procedures used for the classification and the appropriateness and accuracy of the classification should be documented.

Comment: Empirical evidence and/or expert judgment are used to classify items according to categories of the test specifications. For example, professional panels may be used for classifying the items or for determining the appropriateness of the developer's classification scheme. The panel and procedures used should be chosen with care as they will affect the accuracy of the classification.

Standard 3.8

When item tryouts or field tests are conducted, the procedures used to select the sample(s) of test takers for item tryouts and the resulting characteristics of the sample(s) should be documented. When appropriate, the sample(s) should be as representative as possible of the population(s) for which the test is intended.

Comment: Conditions which may differentially affect performance on the test items by the sample(s) as compared to the intended population(s) should be documented when appropriate. As an example, test takers may be less motivated when they know their scores will not have an impact on them.

Standard 3.9

When a test developer evaluates the psychometric properties of items, the classical or item response theory (IRT) model used for evaluating the psychometric properties of items should be documented. The sample used for estimating item properties should be de-

scribed and should be of adequate size and diversity for the procedure. The process by which items are selected and the data used for item selection, such as item difficulty, item discrimination, and/or item information, should also be documented. When IRT is used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.

Comment: Although overall sample size is important, it is important also that there be an adequate number of cases in regions critical to the determination of the psychometric properties of items. If the test is to achieve greatest precision in a particular part of the score scale and this consideration affects item selection, the manner in which item statistics are used needs to be carefully described. When IRT is used as the basis of test development, it is important to document the adequacy of fit of the model to the data. This is accomplished by providing information about the extent to which IRT assumptions (e.g., unidimensionality, local item independence, or equality of slope parameters) are satisfied.

Test developers should show that any differences between the administration conditions of the field test and the final form do not affect item performance. Conditions that can affect item statistics include item position, time limits, length of test, mode of testing (e.g., paper-and-pencil versus computer-administered), and use of calculators or other tools. For example, in field testing items, those placed at the end of a test might obtain poorer item statistics than those inserted within the test.

Standard 3.10

Test developers should conduct cross-validation studies when items are selected primarily on the basis of empirical relationships rather than on the basis of content or theoretical considerations. The extent to which the different studies identify the same item set should be documented.

Comment: When data-based approaches to test development are used, items are selected primarily on the basis of their empirical relationships with an external criterion, their relationships with one another, or their power to discriminate among groups of individuals. Under these circumstances, it is likely that some items will be selected based on chance occurrences in the data used. Administering the test to a comparable sample of test takers or a hold-out sample provides a means by which the tendency to select items by chance can be determined.

Standard 3.11

Test developers should document the extent to which the content domain of a test represents the defined domain and test specifications.

Comment: Test developers should provide evidence of the extent to which the test items and scoring criteria represent the defined domain. This affords a basis to help determine whether performance on the test can be generalized to the domain that is being assessed. This is especially important for tests that contain a small number of items such as performance assessments. Such evidence may be provided by expert judges.

Standard 3.12

The rationale and supporting evidence for computerized adaptive tests should be documented. This documentation should include procedures used in selecting subsets of items for administration, in determining the starting point and termination conditions for the test, in scoring the test, and for controlling item exposure.

Comment: It is important to assure that documentation of the procedures does not compromise the security of the test items.

If a computerized adaptive test is intended to measure a number of different content subcategories, item selection procedures are to assure that the subcategories are adequately represented by the items presented to the test taker.

STANDARDS

TEST DEVELOPMENT AND REVISION / PART I

Standard 3.13

When a test score is derived from the differential weighting of items, the test developer should document the rationale and process used to develop, review, and assign item weights. When the item weights are obtained based on empirical data, the sample used for obtaining item weights should be sufficiently large and representative of the population for which the test is intended. When the item weights are obtained based on expert judgment, the qualifications of the judges should be documented.

Comment: Changes in the population of test takers, along with other changes such as changes in instructions, training, or job requirements, may impact the original derived item weights, necessitating subsequent studies after an appropriate period of time.

Standard 3.14

The criteria used for scoring test takers' performance on extended-response items should be documented. This documentation is especially important for performance assessments, such as scorable portfolios and essays, where the criteria for scoring may not be obvious to the user.

Comment: The completeness and clarity of the test specifications, including the definition of the domain, are essential in developing the scoring criteria. The test developer needs to provide a clear description of how the test scores are intended to be interpreted to help ensure the appropriateness of the scoring procedures.

Standard 3.15

When using a standardized testing format to collect structured behavior samples, the domain, test design, test specifications, and materials should be documented as for any other test. Such documentation should include a clear definition of the behavior expected of the test takers, the nature of the expected responses, and any materials or directions that are necessary to carry out the testing.

Comment: In developing a prompt, the age, language, experience, and ability level of test takers should be considered, as should other possible unique sources of difficulty for groups in the population to be tested. Test directions that specify time allowances, nature of the responses expected, and rules regarding use of supplementary materials, such as notes, references, dictionaries, calculators, or manipulatives such as lab equipment, may be established via field testing.

Standard 3.16

If a short form of a test is prepared, for example, by reducing the number of items on the original test or organizing portions of a test into a separate form, the specifications of the short form should be as similar as possible to those of the original test. The procedures used for the reduction of items should be documented.

Comment: The extent to which the specifications of the short form differ from those of the original test, and the implications of such differences for interpreting the scores derived from the short form, should be documented.

Standard 3.17

When previous research indicates that irrelevant variance could confound the domain definition underlying the test, then to the extent feasible, the test developer should investigate sources of irrelevant variance. Where possible, such sources of irrelevant variance should be removed or reduced by the test developer.

Standard 3.18

For tests that have time limits, test development research should examine the degree to which scores include a speed component and evaluate the appropriateness of that component, given the domain the test is designed to measure.

Standard 3.19

The directions for test administration should be presented with sufficient clarity and empha-

sis so that it is possible for others to replicate adequately the administration conditions under which the data on reliability and validity, and, where appropriate, norms were obtained.

Comment: Because all people administering tests, including those in schools, industry, and clinics, need to follow test administration conditions carefully, it is essential that test administrators receive detailed instructions on test administration guidelines and procedures.

Standard 3.20

The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample material, practice or sample questions, criteria for scoring, and a representative item identified with each major area in the test's classification or domain should be provided to the test takers prior to the administration of the test or included in the testing material as part of the standard administration instructions.

Comment: For example, in a personality inventory it may be intended that test takers give the first response that occurs to them. Such an expectation should be made clear in the inventory directions. As another example, in directions for interest or occupational inventories, it may be important to specify whether test takers are to mark the activities they would like ideally or whether they are to consider both their opportunity and their ability realistically.

The extent and nature of practice materials and directions depend on expected levels of knowledge among test takers. For example, in using a novel test format, it may be very important to provide the test taker a practice opportunity as part of the test administration. In some testing situations, it may be important for the instructions to address such matters as the effects that guessing and time limits have on test scores. If expansion or elaboration of the test instructions is permitted, the condi-

tions under which this may be done should be stated clearly in the form of general rules and by giving representative examples. If no expansion or elaboration is to be permitted, this should be stated explicitly. Publishers should include guidance for dealing with typical questions from test takers. Users should be instructed how to deal with questions that may arise during the testing period.

Standard 3.21

If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified, and a rationale for permitting the different conditions should be documented.

Comment: In deciding whether the conditions of administration can vary, the test developer needs to consider and study the potential effects of varying conditions of administration. If conditions of administration vary from the conditions studied by the test developer or from those used in the development of norms, the comparability of the test scores may be weakened and the applicability of the norms can be questioned.

Standard 3.22

Procedures for scoring and, if relevant, scoring criteria should be presented by the test developer in sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical if tests can be scored locally.

Standard 3.23

The process for selecting, training, and qualifying scorers should be documented by the test developer. The training materials, such as the

STANDARDS

TEST DEVELOPMENT AND REVISION / PART I

scoring rubrics and examples of test takers' responses that illustrate the levels on the score scale, and the procedures for training scorers should result in a degree of agreement among scorers that allows for the scores to be interpreted as originally intended by the test developer. Scorer reliability and potential drift over time in raters' scoring standards should be evaluated and reported by the person(s) responsible for conducting the training session.

Standard 3.24

When scoring is done locally and requires scorer judgment, the test user is responsible for providing adequate training and instruction to the scorers and for examining scorer agreement and accuracy. The test developer should document the expected level of scorer agreement and accuracy.

Comment: A common practice of test developers is to provide examples of training materials (e.g., scoring rubrics, test takers' responses at each score level) and procedures when scoring is done locally and requires scorer judgment.

Standard 3.25

A test should be amended or revised when new research data, significant changes in the domain represented, or newly recommended conditions of test use may lower the validity of test score interpretations. Although a test that remains useful need not be withdrawn or revised simply because of the passage of time, test developers and test publishers are responsible for monitoring changing conditions and for amending, revising, or withdrawing the test as indicated.

Comment: Test developers need to consider a number of factors that may warrant the revision of a test, including outdated test content and language. If an older version of a test is used when a newer version has been published or made available, test users are responsible for

providing evidence that the older version is as appropriate as the new version for that particular test use.

Standard 3.26

Tests should be labeled or advertised as "revised" only when they have been revised in significant ways. A phrase such as "with *minor modification*" should be used when the test has been modified in minor ways. The score scale should be adjusted to account for these modifications, and users should be informed of the adjustments made to the score scale.

Comment: It is the test developer's responsibility to determine whether revisions to a test would influence test score interpretations. If test score interpretations would be affected by the revisions, it would then be appropriate to label the test "revised." When tests are revised, the nature of the revisions and their implications on test score interpretations should be documented.

Standard 3.27

If a test or part of a test is intended for research use only and is not distributed for operational use, statements to this effect should be displayed prominently on all relevant test administration and interpretation materials that are provided to the test user.

Comment: This standard refers to tests that are intended for research use only and does not refer to standard test development functions that occur prior to the operational use of a test (e.g., field testing).

4. SCALES, NORMS, AND SCORE COMPARABILITY

Background

Test scores are reported on scales designed to assist score interpretation. Typically, scoring begins with responses to separate test items, which are often coded using 0 or 1 to represent wrong/right or negative/positive, but sometimes using numerical values to indicate finer response gradations. Then the item scores are combined, often by addition but sometimes by a more elaborate procedure, to obtain a *raw score*. Raw scores are determined, in part, by features of a test such as test length, choice of time limit, item difficulties, and the circumstances under which the test is administered. This makes raw scores difficult to interpret in the absence of further information. Interpretation and statistical analyses may be facilitated by converting raw scores into an entirely different set of values called *derived scores* or *scale scores*. The various scales used for reporting scores on college admissions tests, the standard scores often used to report results for intelligence scales or vocational interest and personality inventories, and the *grade equivalents* reported for achievement tests in the elementary grades are examples of scale scores. The process of developing such a score scale is called *scaling* a test. Scale scores may aid interpretation by indicating how a given score compares to those of other test takers, by enhancing the comparability of scores obtained using different forms of a test, or in other ways.

Another way of assisting score interpretation is to establish *standards* or *cut scores* that distinguish different score ranges. In some cases, a single cut score may define the boundary between passing and failing. In other cases, a series of cut scores may define distinct proficiency levels. Cut scores may be established for either raw or scale scores. Both scale scores and standards or cut scores can be central to the use and interpretation of test scores. For

that reason, their defensibility is an important consideration in test validation. There is a close connection between standards or cut scores and certain scale scores. If the successive score ranges defined by a series of cut scores are relabeled, say 0, 1, 2, and so on, then a scale score has been created.

In addition to facilitating interpretations of a single test form considered in isolation, scale scores are often created to enhance comparability across different forms of the same test, across test formats or administration conditions, or even across tests designed to measure different constructs (e.g., related subtests in a battery). Equated scores from alternate forms of a test can often be interpreted more easily when expressed in scale score units rather than raw score units. Scaling may be used to place scores from different levels of an achievement test on a continuous scale and thereby facilitate inferences about growth or development. Scaling can also enhance the comparability of scores derived from tests in different areas, as in subtests within an aptitude, interest, or achievement battery.

Norm-Referenced and Criterion-Referenced Score Interpretations

Individual raw scores or scale scores are often referred to the distribution of scores for one or more comparison groups to draw useful inferences about an individual's performance. Test score interpretations based on such comparisons are said to be norm-referenced. Percentile rank norms, for example, indicate the standing of an individual or group within a defined population of individuals or groups. An example of such a comparison group might be fourth-grade students in the United States, tested in the last 2 months of a recent school year. Percentiles, averages, or other statistics for such reference groups are called norms. By showing

how the test score of a given examinee compares to those of others, norms assist in the classification or description of examinees.

Other test score interpretations make no direct reference to the performance of other examinees. These interpretations may take a variety of forms; most are collectively referred to as *criterion-referenced* interpretations. Derived scores supporting such interpretations may indicate the likely proportion of correct responses on some larger domain of items, or the probability of an examinee's answering particular sorts of items correctly. Other *criterion-referenced* interpretations may indicate the likelihood that some psychopathology is present. Still other *criterion-referenced* interpretations indicate the probability that an examinee's level of tested knowledge or skill is adequate to perform successfully in some other setting; such probabilities may be summarized in an expectancy table. Scale scores to support such *criterion-referenced* score interpretations are often developed on the basis of statistical analyses of the relationships of test scores to other variables.

Some scale scores are developed primarily to support *norm-referenced* interpretations and others, *criterion-referenced* interpretations. In practice, however, there is not always a sharp distinction. Both *criterion-referenced* and *norm-referenced* scales may be developed and used for the same test scores. Moreover, a *norm-referenced* score scale originally developed, for example, to indicate performance relative to some specific reference population might, over time, also come to support *criterion-referenced* interpretations. This could happen as research and experience brought increased understanding of the capabilities implied by different scale score levels. Conversely, results of an educational assessment might be reported on a scale consisting of several ordered proficiency levels, defined by descriptions of the kinds of tasks students at each level were able to perform. That would be a *criterion-referenced* scale, but once the

distribution of scores over levels was reported, say, for all eighth-grade students in a given state, individual students' scores would also convey information about their standing relative to that tested population.

Interpretations based on cut scores may likewise be either *criterion-referenced* or *norm-referenced*. If qualitatively different descriptions are attached to successive score ranges, a *criterion-referenced* interpretation is supported. For example, the descriptions of performance levels in some assessment task scoring rubrics can enhance score interpretation by summarizing the capabilities that must be demonstrated to merit a given score. In other cases, *criterion-referenced* interpretations may be based on empirically determined relationships between test scores and other variables. But when tests are used for selection, it may be appropriate to rank-order examinees according to their test performance and establish a cut score so as to select a prespecified number or proportion of examinees from one end of the distribution, if the selection use is otherwise supported by relevant reliability and validity evidence. In such cases, the cut score interpretation is *norm-referenced*; the labels *reject* or *fail* versus *accept* or *pass* are determined solely by an examinee's standing relative to others tested.

Criterion-referenced interpretations based on cut scores are sometimes criticized on the grounds that there is very rarely a sharp distinction of any kind between those just below versus just above a cut score. A neuropsychological test may be helpful in diagnosing some particular impairment, for example, but the probability that the impairment is present is likely to increase continuously as a function of the test score. Cut scores may nonetheless aid in formulating rules for reaching decisions on the basis of test performance. It should be recognized, however, that the probability of misclassification will generally be relatively high for persons with scores close to the cut points.

PART I / SCALES, NORMS, AND SCORE COMPARABILITY

Norms

The validity of norm-referenced interpretations depends in part on the appropriateness of the reference group to which test scores are compared. Norms based on hospitalized patients, for example, might be inappropriate for some interpretations of nonhospitalized patients' scores. Thus, it is important that reference populations be carefully defined and clearly described. Validity of such interpretations also depends on the accuracy with which norms summarize the performance of the reference population. That population may be small enough that essentially the entire population can be tested (e.g., all pupils at a given grade level in a given district tested on the same occasion). Often, however, only a sample of examinees from the reference population is tested. It is then important that the norms be based on a technically sound, representative, scientific sample of sufficient size. Patients in a few hospitals in a small geographic region are unlikely to be representative of all patients in the United States, for example. Moreover, the appropriateness of norms based on a given sample may diminish over time. Thus, for tests that have been in use for a number of years, periodic review is generally required to assure the continued utility of norms. Renorming may be required to maintain the validity of norm-referenced test score interpretations.

More than one reference population may be appropriate for the same test. For example, achievement test performance might be interpreted by reference to local norms based on sampling from a particular school district, norms for a state or type of community, or national norms. For other tests, norms might be based on occupational or educational classifications. Descriptive statistics for all examinees who happen to be tested during a given period of time (sometimes called *user norms* or *program norms*) may be useful for some purposes, such as describing trends over time. But there must be sound reason to regard that

group of test takers as an appropriate basis for such inferences. When there is a suitable rationale for using such a group, the descriptive statistics should be clearly characterized as being based on a sample of persons routinely tested as part of an ongoing program.

Comparability and Equating

Many test uses involve different versions of the same test, which yield scores that can be used interchangeably even though they are based on different sets of items. In testing programs that offer a choice of examination dates, for example, test security may be compromised if the same form is used repeatedly. Other testing applications may entail repeated measurements of the same individuals, perhaps to measure change in levels of psychological dysfunction, change in attitudes, or educational progress. In such contexts, reuse of the same set of test items may result in correlated errors of measurement and biased estimates of change. When distinct forms of a test are constructed to the same explicit content and statistical specifications and administered under identical conditions, they are referred to as *alternate forms* or sometimes *parallel or equivalent* forms. The process of placing scores from such alternate forms on a common scale is called *equating*. Equating is analogous to the calibration of different balances so that they all indicate the same weight for any given object. However, the equating process for test scores is more complex. It involves small statistical adjustments to account for minor differences in the difficulty and statistical properties of the alternate forms.

In theory, equating should provide accurate score conversions for any set of persons drawn from the examinee population for which the test is designed. Furthermore, the same score conversion should be appropriate regardless of the score interpretation or use intended. It is not possible to construct conversions with these ideal properties between scores on

tests that measure different constructs; that differ materially in difficulty, reliability, time limits, or other conditions of administration; or that are designed to different specifications.

There is another assessment approach that may provide interchangeable scores based on responses to different items using different methods, not referred to as equating. This is the use of *adaptive tests*. It has long been recognized that little is learned from examinees' responses to items that are much too easy or much too difficult for them. Consequently, some testing procedures use only a subset of the available items with each examinee in order to avoid boredom or frustration, or to shorten testing time. An adaptive test consists of a pool of items together with rules for selecting a subset of those items to be administered to an individual examinee, and a procedure for placing different examinees' scores on a common scale. The selection of successive items is based in part on the examinee's responses to previous items. The item pool and item selection rules may be designed so that each examinee receives a representative set of items, of appropriate difficulty. The selection rules generally assure that an acceptable degree of precision is attained before testing is terminated. At one time, such tailored testing was limited to certain individually administered psychological tests. With advances in item response theory (IRT) and in computer technology, however, adaptive testing is becoming more sophisticated. With some adaptive tests, it may happen that two examinees rarely if ever respond to precisely the same set of items. Moreover, two examinees taking the same adaptive test may be given sets of items that differ markedly in difficulty. Nevertheless, when certain statistical and content conditions are met, test scores produced by an adaptive testing system can function like scores from equated alternate forms.

Scaling to Achieve Comparability

The term *equating* is properly reserved only for score conversions derived for alternate forms of the same test. It is often useful, however, to compare scores from tests that cannot, in theory, be equated. For example, it may be desirable to interpret scores from a shortened (and hence less reliable) form of a test by first converting them to corresponding scores on the full-length version. For the evaluation of examinee growth over time, it may be desirable to develop scales that span a broad range of developmental or educational levels. Test revision often brings a need for some linkage between scores obtained using newer and older editions. International comparative studies or use with hearing-impaired examinees may require test forms in different languages. In still other cases, linkages or alignments may be created between tests measuring different constructs, perhaps comparing an aptitude with a form of behavior, or linking measures of achievement in several content areas. Scores from such tests may sometimes be aligned or presented in a concordance table to aid users in estimating relative performance on one test from performance on another.

Score conversions to facilitate such comparisons may be described using terms like linkage, calibration, concordance, projection, moderation, or anchoring. These weaker score linkages may be technically sound and may fully satisfy desired goals of comparability for one purpose or for one subgroup of examinees, but they cannot be assumed to be stable over time or invariant across multiple subgroups of the examinee population nor is there any assurance that scores obtained using different tests will be equally accurate. Thus, their use for other purposes or with other populations than originally intended may require additional research. For example, a score conversion that was accurate for a group of native speakers might systematically overpredict or underpredict the scores of a group of nonnative speakers.

Cut Scores

A critical step in the development and use of some tests is to establish one or more cut points dividing the score range to partition the distribution of scores into categories. These categories may be used just for descriptive purposes or may be used to distinguish among examinees for whom different programs are deemed desirable or different predictions are warranted. An employer may determine a cut score to screen potential employees or promote current employees; a school may use test scores to decide which of several alternative instructional programs would be most beneficial for a student; in granting a professional license, a state may specify a minimum passing score on a licensure test.

These examples differ in important respects, but all involve delineating categories of examinees on the basis of test scores. Such cut scores embody the rules according to which tests are used or interpreted. Thus, in some situations the validity of test interpretations may hinge on the cut scores. There can be no single method for determining cut scores for all tests or for all purposes, nor can there be any single set of procedures for establishing their defensibility. These examples serve only as illustrations.

The first example, that of an employer hiring all those who earn scores above a given level on an employment test, is most straightforward. Assuming that the employment test is valid for its intended use, average job performance would typically be expected to rise steadily, albeit slowly, with each increment in test score, at least for some range of scores surrounding the cut point. In such a case the designation of the particular value for the cut point may be largely determined by the number of persons to be hired or promoted. There is no sharp difference between those just below the cut point and those just above it, and the use of the cut score does not entail any criterion-referenced interpretation. This method

of establishing a cut score may be subject to legal requirements with respect to the nature of the validity and reliability evidence needed to support the use of rank-order selections and the unavailability of effective alternative selection methods, if it has a disproportionate effect on one or more subgroups of employees or prospective employees.

In the second example, a school district might structure its courses in writing around *three categories of needs*. For children whose proficiency is least developed, instruction might be provided in small groups, with considerable individual attention to assist them in creating meaningful written stories grounded in their own experience. For children whose proficiency was further developed, more emphasis might be placed on systematic exploration of the stages of the writing process. Instruction for children at the highest proficiency level might emphasize mastery of specific writing genres or prose structures used in more formal writing. In an appropriate implementation of such a program, children could easily be transferred from one level to another if their original placement was in error or as their proficiency increased. Ideally, cut scores delineating categories in this application would be based on research demonstrating empirically that pupils in successive score ranges did most often benefit more from the respective treatments to which they were assigned than from the alternatives available. It would typically be found that between those score ranges in which one or another instructional treatment was clearly superior, there was an intermediate region in which neither treatment was clearly preferred. The cut score might be located somewhere in that intermediate region.

In the final example, that of a professional licensure examination, the cut score represents an informed judgment that those scoring below it are likely to make serious errors for want of the knowledge or skills tested. Little evidence apart from errors made on the test itself may document the need to deny the right to prac-

STANDARDS

SCALES, NORMS, AND SCORE COMPARABILITY / PART I

rice the profession. No test is perfect, of course, and regardless of the cut score chosen, some examinees with inadequate skills are likely to pass and some with adequate skills are likely to fail. The relative probabilities of such false positive and false negative errors will vary depending on the cut score chosen. A given probability of exposing the public to potential harm by issuing a license to an incompetent individual (false positive) must be weighed against some corresponding probability of denying a license to, and thereby disenfranchising, a qualified examinee (false negative). Changing the cut score to reduce either probability will increase the other, although both kinds of errors can be minimized through sound test design that anticipates the role of the cut score in test use and interpretation. Determining cut scores in such situations cannot be a purely technical matter, although empirical studies and statistical models can be of great value in informing the process.

Cut scores embody value judgments as well as technical and empirical considerations. Where the results of the standard-setting process have highly significant consequences, and especially where large numbers of examinees are involved, those responsible for establishing cut scores should be concerned that the process by which cut scores are determined be clearly documented and defensible. The qualifications of any judges involved in standard setting and the process by which they are selected are part of that documentation. Care must be taken to assure that judges understand what they are to do. The process must be such that well-qualified judges can apply their knowledge and experience to reach meaningful and relevant judgments that accurately reflect their understandings and intentions. A sufficiently large and representative group of judges should be involved to provide reasonable assurance that results would not vary greatly if the process were replicated.

Standard 4.1

Test documents should provide test users with clear explanations of the meaning and intended interpretation of derived score scales, as well as their limitations.

Comment: All scales (raw score or derived) may be subject to misinterpretation. Sometimes scales are extrapolated beyond the range of available data or are interpolated without sufficient data points. Grade- and age-equivalent scores have been criticized in this regard, but percentile ranks and standard score scales are also subject to misinterpretation. If the nature or intended uses of a scale are novel, it is especially important that its uses, interpretations, and limitations be clearly described. Illustrations of appropriate versus inappropriate interpretations may be helpful, especially for types of scales or interpretations that may be unfamiliar to most users. This standard pertains to score scales intended for criterion-referenced as well as for norm-referenced interpretation.

Standard 4.2

The construction of scales used for reporting scores should be described clearly in test documents.

Comment: When scales, norms, or other interpretive systems are provided by the test developer, technical documentation should enable users to judge the quality and precision of the resulting derived scores. This standard pertains to score scales intended for criterion-referenced as well as for norm-referenced interpretation.

Standard 4.3

If there is sound reason to believe that specific misinterpretations of a score scale are likely, test users should be explicitly forewarned.

Comment: Test publishers and users can reduce misinterpretations of grade-equivalent scores, for example, by ensuring that such scores are accompanied by instructions that make clear that grade-equivalent scores do not represent a standard of growth per year or grade and that roughly 50% of the students tested in the standardization sample should by definition fall below grade level. As another example, a score scale point originally defined as the mean of some reference population should no longer be interpreted as representing average performance if the scale is held constant over time and the examinee population changes.

Standard 4.4

When raw scores are intended to be directly interpretable, their meanings, intended interpretations, and limitations should be described and justified in the same manner as is done for derived score scales.

Comment: In some cases the items in a test are a representative sample of a well-defined domain of items. The proportion correct on the test may then be interpreted as an estimate of the proportion of items in the domain that could be answered correctly. In other cases, different interpretations may be attached to scores above or below one or another cut score. Support should be offered for any such interpretations recommended by the test developer.

Standard 4.5

Norms, if used, should refer to clearly described populations. These populations should include individuals or groups to whom test users will ordinarily wish to compare their own examinees.

Comment: It is the responsibility of test developers to describe norms clearly and the responsibility of test users to employ norms appropriately. Users need to know the applicability of a test to different groups. Differentiated norms or sum-

mary information about differences between gender, ethnic, language, disability, grade, or age groups, for example, may be useful in some cases. The permissible uses of such differentiated norms and related information may be limited by law. Users also need to be made alert to situations in which norms are less appropriate for some groups or individuals than others. On an occupational interest inventory, for example, norms for persons actually engaged in an occupation may be inappropriate for interpreting the scores of persons not so engaged. As another example, the appropriateness of norms for personality inventories or relationship scales may differ depending upon an examinee's sexual orientation.

Standard 4.6

Reports of norming studies should include precise specification of the population that was sampled, sampling procedures and participation rates, any weighting of the sample, the dates of testing, and descriptive statistics. The information provided should be sufficient to enable users to judge the appropriateness of the norms for interpreting the scores of local examinees. Technical documentation should indicate the precision of the norms themselves.

Comment: Scientific sampling is important if norms are to be representative of intended populations. For example, schools already using a given published test and volunteering to participate in a norming study for that test should not be assumed to be representative of schools in general. In addition to sampling procedures, participation rates should be reported, and the method of calculating participation rates should be clearly described. Studies that are designed to be nationally representative often use weights so that the weighted sample better represents the nation than does the unweighted sample. When weights are used, it is important that the procedure for deriving the weights be described and that the demographic representa-

tion of both the weighted and the unweighted samples be given. If norming data are collected under conditions in which student motivation in completing the test is likely to differ from that expected during operational use, this should be clearly documented. Likewise, if the instructional histories of students in the norming sample differ systematically from those to be expected during operational test use, that fact should be noted. Norms based on samples cannot be perfectly precise. Even though the imprecision of norm-referenced interpretations due to imperfections in the norms themselves may be small compared to that due to measurement error, estimates of the precision of norms should be available in technical documentation. For example, standard errors based on the sample design might be presented. In some testing applications, norms based on all examinees tested over a given period of time may be useful for some purposes. Such norms should be clearly characterized as being based on a sample of persons routinely tested as part of an ongoing testing program.

Standard 4.7

If local examinee groups differ materially from the populations to which norms refer, a user who reports derived scores based on the published norms has the responsibility to describe such differences if they bear upon the interpretation of the reported scores.

Comment: In employment settings, the qualifications of local examinee groups may fluctuate depending on recruitment or referral procedures as well as market conditions. In such cases, appropriate test use and interpretation may not require documentation or cautions concerning departures from characteristics of the norming population.

Standard 4.8

When norms are used to characterize examinee groups, the statistics used to summarize

each group's performance and the norms to which those statistics are referred should be clearly defined and should support the intended use or interpretation.

Comment: Group means are distributed differently from individual scores. For example, it is not possible to determine the percentile rank of a school's average test score if all that is known are the percentile ranks of each of that school's students. It may sometimes be useful to develop special norms for group means, but when the sizes of the groups differ materially or when some groups are much more heterogeneous than others, the construction and interpretation of group norms is problematical. One common and acceptable procedure is to report the percentile rank of the median group member, for example, the median percentile rank of the pupils tested in a given school.

Standard 4.9

When raw score or derived score scales are designed for criterion-referenced interpretation, including the classification of examinees into separate categories, the rationale for recommended score interpretations should be clearly explained.

Comment: Criterion-referenced interpretations are score-based descriptions or inferences that do not take the form of comparisons to the test performance of other examinees. Examples include statements that some psychopathology is likely present, that a prospective employee possesses specific skills required in a given position, or that a child scoring above a certain score point can successfully apply a given set of skills. Such interpretations may refer to the absolute levels of test scores or to patterns of scores for an individual examinee. Whenever the test developer recommends such interpretations, the rationale and empirical basis should be clearly presented. Serious efforts should be made whenever possible to obtain independent

evidence concerning the soundness of such score interpretations. Criterion-referenced and norm-referenced scales are not mutually exclusive. Given adequate supporting data, scores may be interpreted by both approaches, not necessarily just one or the other.

Standard 4.10

A clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably. In some cases, direct evidence of score equivalence may be provided. In other cases, evidence may come from a demonstration that the theoretical assumptions underlying procedures for establishing score comparability have been sufficiently satisfied. The specific rationale and the evidence required will depend in part on the intended uses for which score equivalence is claimed.

Comment: Support should be provided for any assertion that scores obtained using different items or testing materials, or different testing procedures, are interchangeable for some purpose. This standard applies, for example, to alternate forms of a paper-and-pencil test or to alternate sets of items taken by different examinees in computerized adaptive testing. It also applies to test forms administered in different formats (e.g., paper-and-pencil and computerized tests) or test forms designed for individual versus group administration. Score equivalence is easiest to establish when different forms are constructed following identical procedures and then equated statistically. When that is not possible, for example, in cases where different test formats are used, additional evidence may be required to establish the requisite degree of score equivalence for the intended context and purpose. When recommended inferences or actions are based solely on classifications of examinees into one of two or more categories, the rationale and evidence should address consistency of classification. If the only

score reported and used is a pass-fail decision, for example, then the form-to-form equivalence of measurements for examinees far above or far below the cut score is of no concern. Some testing accommodations may only affect the dependence of test scores on capabilities irrelevant to the construct the test is intended to measure. Use of a large-print edition, for example, assures that performance does not depend on the ability to perceive standard-size print. In such cases, relatively modest studies or professional judgment may be sufficient to support claims of score equivalence.

Standard 4.11

When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions or other linkages were established and on the accuracy of equating functions.

Comment: The fundamental concern is to show that equated scores measure essentially the same construct, with very similar levels of reliability and conditional standard errors of measurement. Technical information should include the design of equating studies, the statistical methods used, the size and relevant characteristics of examinee samples used in equating studies, and the characteristics of any anchor tests or linking items. Standard errors of equating functions should be estimated and reported whenever possible. Sample sizes permitting, it may be informative to determine equating functions independently for identifiable subgroups of examinees. It may also be informative to use two anchor forms and to conduct the equating using each of the anchors. In some cases, equating functions may be determined independently using different statistical methods. The correspondence of separate functions obtained by such methods can lend support to the adequacy of the equating results. Any substantial disparities found by such methods

STANDARDS

should be resolved or reported. To be most useful, equating error should be presented in units of the reported score scale. For testing programs with cut scores, equating error near the cut score is of primary importance. The degree of scrutiny of equating functions should be commensurate with the extent of test use anticipated and the importance of the decisions the test scores are intended to inform.

Standard 4.12

In equating studies that rely on the statistical equivalence of examinee groups receiving different forms, methods of assuring such equivalence should be described in detail.

Comment: Certain equating designs rely on the random equivalence of groups receiving different forms. Often, one way to assure such equivalence is to systematically mix different test forms and then distribute them in a random fashion so that roughly equal numbers of examinees in each group tested receive each form.

Standard 4.13

In equating studies that employ an anchor test design, the characteristics of the anchor test and its similarity to the forms being equated should be presented, including both content specifications and empirically determined relationships among test scores. If anchor items are used, as in some IRT-based and classical equating studies, the representativeness and psychometric characteristics of anchor items should be presented.

Comment: Tests or test forms may be linked via common items embedded within each of them, or a common test administered together with each of them. These common items or tests are referred to as linking items, anchor items, or anchor tests. With such methods, the quality of the resulting equating depends strongly on the adequacy of the anchor tests or items used.

Standard 4.14

When score conversions or comparison procedures are used to relate scores on tests or test forms that are not closely parallel, the construction, intended interpretation, and limitations of those conversions or comparisons should be clearly described.

Comment: Various score conversions or concordance tables have been constructed relating tests at different levels of difficulty, relating earlier to revised forms of published tests, creating score concordances between different tests of similar or different constructs, or for other purposes. Such conversions are often useful, but they may also be subject to misinterpretation. The limitations of such conversions should be clearly described.

Standard 4.15

When additional test forms are created by taking a subset of the items in an existing test form or by rearranging its items and there is sound reason to believe that scores on these forms may be influenced by item context effects, evidence should be provided that there is no undue distortion of norms for the different versions or of score linkages between them.

Comment: Some tests and test batteries are published in both a full-length version and a survey or short version. In other cases, multiple versions of a single test form may be created by rearranging its items. It should not be assumed that performance data derived from the administration of items as part of the initial version can be used to approximate norms or construct conversion tables for alternative intact tests. Due caution is required in cases where context effects are likely, including speeded tests, long tests where fatigue may be a factor, and so on. In many cases, adequate psychometric data may only be obtainable from independent administrations of the alternate forms.

STANDARDS

Standard 4.16

If test specifications are changed from one version of a test to a subsequent version, such changes should be identified in the test manual, and an indication should be given that converted scores for the two versions may not be strictly equivalent. When substantial changes in test specifications occur, either scores should be reported on a new scale or a clear statement should be provided to alert users that the scores are not directly comparable with those on earlier versions of the test.

Comment: Major shifts sometimes occur in the specifications of tests that are used for substantial periods of time. Often such changes take advantage of improvements in item types or of shifts in content that have been shown to improve validity and, therefore, are highly desirable. It is important to recognize, however, that such shifts will result in scores that cannot be made strictly interchangeable with scores on an earlier form of the test.

Standard 4.17

Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which scores are reported.

Comment: In some testing programs, items are introduced into and retired from item pools on an ongoing basis. In other cases, the items in successive test forms may overlap very little, or not at all. In either case, if a fixed scale is used for reporting, it is important to assure that the meaning of the scaled scores does not change over time.

Standard 4.18

If a publisher provides norms for use in test score interpretation, then so long as the test remains in print, it is the publisher's responsibility to assure that the test is renormed with sufficient frequency to permit continued accurate and appropriate score interpretations.

Comment: Test publishers should assure that up-to-date norms are readily available, but it remains the test user's responsibility to avoid inappropriate use of norms that are out of date and to strive to assure accurate and appropriate test interpretations.

Standard 4.19

When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented.

Comment: Cut scores may be established to select a specified number of examinees (e.g., to fill existing vacancies), in which case little further documentation may be needed concerning the specific question of how the cut scores are established, though attention should be paid to legal requirements that may apply. In other cases, however, cut scores may be used to classify examinees into distinct categories (e.g., diagnostic categories, or passing versus failing) for which there are no preestablished quotas. In these cases, the standard-setting method must be clearly documented. Ideally, the role of cut scores in test use and interpretation is taken into account during test design. Adequate precision in regions of score scales where cut points are established is prerequisite to reliable classification of examinees into categories. If standard setting employs data on the score distributions for criterion groups or on the relation of test scores to one or more criterion variables, those data should be summarized in technical documentation. If a judgmental standard-setting process is followed, the method employed should be clearly described, and the precise nature of the judgments called for should be presented, whether those are judgments of persons, of item or test performances, or of other criterion performances predicted by test scores. Documentation should also include the selection and qualification of judges, training provided, any feedback to judges concerning the implications of their provisional judgments.

STANDARDS

SCALES, NORMS, AND SCORE COMPARABILITY / PART I

and any opportunities for judges to confer with one another. Where applicable, variability over judges should be reported. Whenever feasible, an estimate should be provided of the amount of variation in cut scores that might be expected if the standard-setting procedure were replicated.

Standard 4.20

When feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria.

Comment: In employment settings, although it is important to establish that test scores are related to job performance, the precise relation of test and criterion may have little bearing on the choice of a cut score. However, in contexts where distinct interpretations are applied to different score categories, the empirical relation of test to criterion assumes greater importance. Cut scores used in interpreting diagnostic tests may be established on the basis of empirically determined score distributions for criterion groups. With achievement or proficiency tests, such as those used in licensure, suitable criterion groups (e.g., successful versus unsuccessful practitioners) are often unavailable. Nonetheless, it is highly desirable, when appropriate and feasible, to investigate the relation between test scores and performance in relevant practical settings. Note that a carefully designed and implemented procedure based solely on judgments of content relevance and item difficulty may be preferable to an empirical study with an inadequate criterion measure or other deficiencies. Professional judgment is required to determine an appropriate standard-setting approach (or combination of approaches) in any given situation. In general, one would not expect to find a sharp difference in levels of the criterion variable between those just

below versus just above the cut score, but evidence should be provided where feasible of a relationship between test and criterion performance over a score interval that includes or approaches the cut score.

Standard 4.21

When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way.

Comment: Cut scores are sometimes based on judgments about the adequacy of item or test performances (e.g., essay responses to a writing prompt) or performance levels (e.g., the level that would characterize a borderline examinee). The procedures used to elicit such judgments should result in reasonable, defensible standards that accurately reflect the judges' values and intentions. Reaching such judgments may be most straightforward when judges are asked to consider kinds of performances with which they are familiar and for which they have formed clear conceptions of adequacy or quality. When the responses elicited by a test neither sample nor closely simulate the use of tested knowledge or skills in the actual criterion domain, judges are not likely to approach the task with such clear understandings. Special care must then be taken to assure that judges have a sound basis for making the judgments requested. Thorough familiarity with descriptions of different proficiency categories, practice in judging task difficulty with feedback on accuracy, the experience of actually taking a form of the test, feedback on the failure rates entailed by provisional standards, and other forms of information may be beneficial in helping judges to reach sound and principled decisions.

5. TEST ADMINISTRATION, SCORING, AND REPORTING

Background

The usefulness and interpretability of test scores require that a test be administered and scored according to the developer's instructions. When directions to examinees, testing conditions, and scoring procedures follow the same detailed procedures, the test is said to be standardized. Without such standardization, the accuracy and comparability of score interpretations would be reduced. For tests designed to assess the examinee's knowledge, skills, or abilities, standardization helps to ensure that all examinees have the same opportunity to demonstrate their competencies. Maintaining test security also helps to ensure that no one has an unfair advantage.

Occasionally, however, situations arise in which modifications of standardized procedures may be advisable or legally mandated. Persons of different backgrounds, ages, or familiarity with testing may need nonstandard modes of test administration or a more comprehensive orientation to the testing process, in order that all test takers can come to the same understanding of the task. Standardized modes of presenting information or of responding may not be suitable for specific individuals, such as persons with some kinds of disability, or persons with limited proficiency in the language of the test, so that accommodations may be needed (see chapters 9 and 10). Large-scale testing programs generally have established specific procedures to be used in considering and granting accommodations. Some test users feel that any accommodation not specifically required by law could lead to a charge of unfair treatment and discrimination. Although accommodations are made with the intent of maintaining score comparability, the extent to which that is possible may not be known. Comparability of scores may be compromised, and the test may then not measure the same constructs for all test takers.

Tests and assessments differ in their degree of standardization. In many instances different examinees are given not the same test form, but equivalent forms that have been shown to yield comparable scores. Some assessments permit examinees to choose which tasks to perform or which pieces of their work are to be evaluated. A degree of standardization can be maintained by specifying the conditions of the choice and the criteria of evaluation of the products. When an assessment permits a certain kind of collaboration, the limits of that collaboration can be specified. With some assessments, test administrators may be expected to tailor their instructions to help assure that all examinees understand what is expected of them. In all such cases, the goal remains the same: to provide accurate and comparable measurement for everyone, and unfair advantage to no one. The degree of standardization is dictated by that goal, and by the intended use of the test.

Standardized directions to test takers help to ensure that all test takers understand the mechanics of test taking. Directions generally inform test takers how to make their responses, what kind of help they may legitimately be given if they do not understand the question or task, how they can correct inadvertent responses, and the nature of any time constraints. General advice is sometimes given about omitting item responses. Many tests, including computer-administered tests, require special equipment. Practice exercises are often presented in such cases to ensure that the test taker understands how to operate the equipment. The principle of standardization includes orienting test takers to materials with which they may not be familiar. Some equipment may be provided at the testing site, such as shop tools or balances. Opportunity for test takers to practice with the equipment will often be appropriate, unless using the equipment is the purpose of the test.

Tests are sometimes administered by computer, with test responses made by keyboard, computer mouse, or similar device. Although many test takers are accustomed to computers, some are not and may need some brief explanation. Even those test takers who use computers will need to know about some details. Special issues arise in managing the testing environment, such as the arrangement of illumination so that light sources do not reflect on the computer screen, possibly interfering with display legibility. Maintaining a quiet environment can be challenging when candidates are tested separately, starting at different times and finishing at different times from neighboring test takers. Those who administer computer-based tests require training in the hardware and software used for the test, so that they can deal with problems that may arise in human-computer interactions.

Standardized scoring procedures help to ensure accurate scoring and reporting, which are essential in all circumstances. When scoring is done by machine, the accuracy of the machine is at issue, including any scoring algorithm. When scoring is done by human judges, scorers require careful training. Regular monitoring can also help to ensure that every test protocol is scored according to the same standardized criteria and that the criteria do not change as the test scorers progress through the submitted test responses.

Test scores, per se, are not readily interpreted without other information, such as norms or standards, indications of measurement error, and descriptions of test content. Just as a temperature of 50° in January is warm for Minnesota and cool for Florida, a test score of 50 is not meaningful without some context. When the scores are to be reported to persons who are not technical specialists, interpretive material can be provided that is readily understandable to those receiving the report. Often, the test user

provides an interpretation of the results for the test taker, suggesting the limitations of the results and the relationship of any reported scores to other information. Scores on some tests are not designed to be released to test takers; only broad test interpretations, or dichotomous classifications, such as pass/fail, are intended to be reported.

Interpretations of test results are sometimes prepared by computer systems. Such interpretations are generally based on a combination of empirical data and expert judgment and experience. In some professional applications of individualized testing, the computer-prepared interpretations are communicated by a professional, possibly with modifications for special circumstances. Such test interpretations require validation. Consistency with interpretations provided by nonalgorithmic approaches is clearly a concern.

In some large-scale assessments, the primary target of assessment is not the individual test taker but is a larger unit, such as a school district or an industrial plant. Often, different test takers are given different sets of items, following a carefully balanced matrix sampling plan, to broaden the range of information that can be obtained in a reasonable time period. The results acquire meaning when aggregated over many individuals taking different samples of items. Such assessments may not furnish enough information to support even minimally valid, reliable scores for individuals, as each individual may take only an incomplete test.

Some further issues of administration and scoring are discussed in chapter 3, "Test Development and Revision."

Standard 5.1

Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer, unless the situation or a test taker's disability dictates that an exception should be made.

Comment: Specifications regarding instructions to test takers, time limits, the form of item presentation or response, and test materials or equipment should be strictly observed. In general, the same procedures should be followed as were used when obtaining the data for scaling and norming the test scores. A test taker with a disabling condition may require special accommodation. Other special circumstances may require some flexibility in administration. Judgments of the suitability of adjustments should be tempered by the consideration that departures from standard procedures may jeopardize the validity of the test score interpretations.

Standard 5.2

Modifications or disruptions of standardized test administration procedures or scoring should be documented.

Comment: Information about the nature of modifications of administration should be maintained in secure data files, so that research studies or case reviews based on test records can take this into account. This includes not only special accommodations for particular test takers, but also disruptions in the testing environment that may affect all test takers in the testing session. A researcher may wish to use only the records based on standardized administration. In other cases, research studies may depend on such information to form groups of respondents. Test users or test sponsors should establish policies concerning who keeps the files and who may have access to the files. Whether the information about

modifications is reported to users of test data, such as admissions officers, depends on different considerations (see chapters 8 and 10). If such reports are made, certain cautions may be appropriate.

Standard 5.3

When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing.

Comment: When large-scale testing programs have established strict procedures to be followed, administrators should not depart from these procedures.

Standard 5.4

The testing environment should furnish reasonable comfort with minimal distractions.

Comment: Noise, disruption in the testing area, extremes of temperature, poor lighting, inadequate work space, illegible materials, and so forth are among the conditions that should be avoided in testing situations. The testing site should be readily accessible. Testing sessions should be monitored where appropriate to assist the test taker when a need arises and to maintain proper administrative procedures. In general, the testing conditions should be equivalent to those that prevailed when norms and other interpretative data were obtained.

Standard 5.5

Instructions to test takers should clearly indicate how to make responses. Instructions should also be given in the use of any equipment likely to be unfamiliar to test takers. Opportunity to practice responding should be given when equipment is involved, unless use of the equipment is being assessed.

STANDARDS

Comment: When electronic calculators are provided for use, examinees may need practice in using the calculator. Examinees may need practice responding with unfamiliar tasks, such as a numeric grid, which is sometimes used with mathematics performance items. In computer-administered tests, the method of responding may be unfamiliar to some test takers. Where possible, the practice responses should be monitored to ensure that the test taker is making acceptable responses. In some performance tests that involve tools or equipment, instructions may be needed for unfamiliar tools, unless accommodating to unfamiliar tools is part of what is being assessed. If a test taker is unable to use the equipment or make the responses, it may be appropriate to consider alternative testing modes.

Standard 5.6

Reasonable efforts should be made to assure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent means.

Comment: In large-scale testing programs where the results may be viewed as having important consequences, efforts to assure score integrity should include, when appropriate and practicable, stipulating requirements for identification, constructing seating charts, assigning test takers to seats, requiring appropriate space between seats, and providing continuous monitoring of the testing process. Test developers should design test materials and procedures to minimize the possibility of cheating. Test administrators should note and report any significant instances of testing irregularity. A local change in the date or time of testing may offer an opportunity for fraud. In general, steps should be taken to minimize the possibility of breaches in test security. In any evaluation of work products (e.g., portfolios) steps should be taken to ensure that the product represents the candidate's own work, and that the amount and kind of assistance provided should be consistent with the intent of

the assessment. Ancillary documentation, such as the date when the work was done, may be useful.

Standard 5.7

Test users have the responsibility of protecting the security of test materials at all times.

Comment: Those who have test materials under their control should, with due consideration of ethical and legal requirements, take all steps necessary to assure that only individuals with a legitimate need for access to test materials are able to obtain such access before the test administration, and afterwards as well, if any part of the test will be reused at a later time. Test users must balance test security with the rights of all test takers and test users. When sensitive test documents are challenged, it may be appropriate to employ an independent third party, using a closely supervised secure procedure to conduct a review of the relevant materials. Such secure procedures are usually preferable to placing tests, manuals, and an examinee's test responses in the public record.

Standard 5.8

Test scoring services should document the procedures that were followed to assure accuracy of scoring. The frequency of scoring errors should be monitored and reported to users of the service on reasonable request. Any systematic source of scoring errors should be corrected.

Comment: Clerical and mechanical errors should be examined. Scoring errors should be minimized and, when they are found, steps should be taken promptly to minimize their recurrence.

Standard 5.9

When test scoring involves human judgment, scoring rubrics should specify criteria for scor-

ing. Adherence to established scoring criteria should be monitored and checked regularly. Monitoring procedures should be documented.

Comment: Human scorers may be provided with scoring rubrics listing acceptable alternative responses, as well as general criteria. Consistency of scoring is often checked by rescoring randomly selected test responses and by rescoring some responses from earlier administrations. Periodic checks of the statistical properties (e.g., means, standard deviations) of scores assigned by individual scorers during a scoring session can provide feedback for the scorers, helping them to maintain scoring standards. Lack of consistent scoring may call for retraining or dismissing some scorers or for reexamining the scoring rubrics.

Standard 5.10

When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used.

Comment: Test users should consult the interpretive material prepared by the test developer or publisher and should revise or supplement the material as necessary to present the local and individual results accurately and clearly. Score precision might be depicted by error bands, or likely score ranges, showing the standard error of measurement.

Standard 5.11

When computer-prepared interpretations of test response protocols are reported, the sources, rationale, and empirical basis for these interpretations should be available, and their limitations should be described.

Comment: Whereas computer-prepared interpretations may be based on expert judgment, the interpretations are of necessity based on accumulated experience and may not be able to take into consideration the context of the individual's circumstances. Computer-prepared interpretations should be used with care in diagnostic settings, because they may not take into account other information about the individual test taker, such as age, gender, education, prior employment, and medical history, that provide context for test results.

Standard 5.12

When group-level information is obtained by aggregating the results of partial tests taken by individuals, validity and reliability should be reported for the level of aggregation at which results are reported. Scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established.

Comment: Large-scale assessments often achieve efficiency by "matrix sampling" of the content domain by asking different test takers different questions. The testing then requires less time from each test taker, while the aggregation of individual results provides for domain coverage that can be adequate for meaningful group- or program-level interpretations, such as schools, or grade levels within a locality or particular subject-matter areas. Because the individual receives only an incomplete test, an individual score would have limited meaning. If individual scores are provided, comparisons between scores obtained by different individuals are based on responses to items that may cover different material. Some degree of calibration among incomplete tests can sometimes be made. Such calibration is essential to the comparisons of individual scores.

STANDARDS

Standard 5.13

Transmission of individually identified test scores to authorized individuals or institutions should be done in a manner that protects the confidential nature of the scores.

Comment: Care is always needed when communicating the scores of identified test takers, regardless of the form of communication. Face-to-face communication, as well as telephone and written communication present well-known problems. Transmission by electronic media, including computer networks and facsimile, presents modern challenges to confidentiality.

Standard 5.14

When a material error is found in test scores or other important information released by a testing organization or other institution, a corrected score report should be distributed as soon as practicable to all known recipients who might otherwise use the erroneous scores as a basis for decision making. The corrected report should be labeled as such.

Comment: A material error is one that could change the interpretation of the test score. Innocuous typographical errors would be excluded. Timeliness is essential for decisions that will be made soon after the test scores are received.

Standard 5.15

When test data about a person are retained, both the test protocol and any written report should also be preserved in some form. Test users should adhere to the policies and record-keeping practice of their professional organizations.

Comment: The protocol may be needed to respond to a possible challenge from a test taker. The protocol would ordinarily be

accompanied by testing materials and test scores. Retention of more detailed records of responses would depend on circumstances and should be covered in a retention policy (see the following standard). Record keeping may be subject to legal and professional requirements. Policy for the release of any test information for other than research purposes is discussed in chapter 8.

Standard 5.16

Organizations that maintain test scores on individuals in data files or in an individual's records should develop a clear set of policy guidelines on the duration of retention of an individual's records, and on the availability, and use over time, of such data.

Comment: In some instances, test scores become obsolete over time, no longer reflecting the current state of the test taker. Outdated scores should generally not be used or made available, except for research purposes. In other cases, test scores obtained in past years can be useful as, for example, in longitudinal assessment. The key issue is the valid use of the information. Score retention and disclosure may be subject to legal and professional requirements.

6. SUPPORTING DOCUMENTATION FOR TESTS

Background

The provision of supporting documents for tests is the primary means by which test developers, publishers, and distributors communicate with test users. These documents are evaluated on the basis of their completeness, accuracy, currency, and clarity and should be available to qualified individuals as appropriate. A test's documentation typically specifies the nature of the test; its intended use; the processes involved in the test's development; technical information related to scoring, interpretation, and evidence of validity and reliability; scaling and norming if appropriate to the instrument; and guidelines for test administration and interpretation. The objective of the documentation is to provide test users with the information needed to make sound judgments about the nature and quality of the test, the resulting scores, and the interpretations based on the test scores. The information may be reported in documents such as test manuals, technical manuals, user's guides, specimen sets, examination kits, directions for test administrators and scorers, or preview materials for test takers.

Test documentation is most effective if it communicates information to multiple user groups. To accommodate the breadth of training of professionals who use tests, separate documents or sections of documents may be written for identifiable categories of users such as practitioners, consultants, administrators, researchers, and educators. For example, the test user who administers the tests and interprets the results needs interpretive information or guidelines. On the other hand, those who are responsible for selecting tests need to be able to judge the technical adequacy of the test. Therefore, some combination of technical manuals, user's guides, test manuals, test supplements, examination kits, or

specimen sets ordinarily is published to provide a potential test user or test reviewer with sufficient information to evaluate the appropriateness and technical adequacy of the test. The types of information presented in these documents typically include a description of the intended test-taking population, stated purpose of the test, test specifications, item formats, scoring procedures, and the test development process. Technical data, such as psychometric indices of the items, reliability and validity evidence, normative data, and cut scores or configural rules including those for computer-generated interpretations of test scores also are summarized.

An essential feature of the documentation for every test is a discussion of the known appropriate and inappropriate uses and interpretations of the test scores. The inclusion of illustrations of score interpretations, as they relate to the test developer's intended applications, also will help users make accurate inferences on the basis of the test scores. When possible, illustrations of improper test uses and inappropriate test score interpretations will help guard against the misuse of the test.

Test documents need to include enough information to allow test users and reviewers to determine the appropriateness of the test for its intended purposes. References to other materials that provide more details about research by the publisher or independent investigators should be cited and should be readily obtainable by the test user or reviewer. This supplemental material can be provided in any of a variety of published or unpublished forms; when demand is likely to be low, it may be maintained in archival form, including electronic storage. Test documentation is useful for all test instruments, including those that are developed exclusively for use within a single organization.

STANDARDS

SUPPORTING DOCUMENTATION FOR TESTS / PART I

In addition to technical documentation, descriptive materials are needed in some settings to inform examinees and other interested parties about the nature and content of the test. The amount and type of information will depend on the particular test and application. For example, in situations requiring informed consent, information should be sufficient to develop a reasoned judgment. Such information should be phrased in nontechnical language and should be as inclusive as is consistent with the use of the test scores. The materials may include a general description and rationale for the test; sample items or complete sample tests; and information about conditions of test administration, confidentiality, and retention of test results. For some applications, however, the true nature and purpose of a test are purposely hidden or disguised to prevent faking or response bias. In these instances, examinees may be motivated to reveal more or less of the characteristics intended to be assessed. Under these circumstances, hiding or disguising the true nature or purpose of the test is acceptable provided this action is consistent with legal principles and ethical standards.

This chapter provides general standards for the preparation and publication of test documentation. The other chapters contain specific standards that will be useful to test developers, publishers, and distributors in the preparation of materials to be included in a test's documentation.

Standard 6.1

Test documents (e.g., test manuals, technical manuals, user's guides, and supplemental material) should be made available to prospective test users and other qualified persons at the time a test is published or released for use.

Comment: The test developer or publisher should judge carefully which information should be included in first editions of the test manual, technical manual, or user's guides and which information can be provided in supplements. For low-volume, unpublished tests, the documentation may be relatively brief. When the developer is also the user, documentation and summaries are still necessary.

Standard 6.2

Test documents should be complete, accurate, and clearly written so that the intended reader can readily understand the content.

Comment: Test documents should provide sufficient detail to permit reviewers and researchers to judge or replicate important analyses published in the test manual. For example, reporting correlation matrices in the test document may allow the test user to judge the data upon which decisions and conclusions were based, or describing in detail the sample and the nature of any factor analyses that were conducted will allow the test user to replicate reported studies.

Standard 6.3

The rationale for the test, recommended uses of the test, support for such uses, and information that assists in score interpretation should be documented. Where particular misuses of a test can be reasonably anticipated, cautions against such misuses should be specified.

Comment: Test publishers make every effort to caution test users against known misuses of

STANDARDS

tests. However, test publishers are not required to anticipate all possible misuses of a test. If publishers do know of persistent test misuse by a test user, extraordinary educational efforts may be appropriate.

Standard 6.4

The population for whom the test is intended and the test specifications should be documented. If applicable, the item pool and scale development procedures should be described in the relevant test manuals. If normative data are provided, the norming population should be described in terms of relevant demographic variables, and the year(s) in which the data were collected should be reported.

Comment: Known limitations of a test for certain populations also should be clearly delineated in the test documents. In addition, if the test is available in more than one language, test documents should provide information on the translation or adaptation procedures, on the demographics of each norming sample, and on score interpretation issues for each language into which the test has been translated.

Standard 6.5

When statistical descriptions and analyses that provide evidence of the reliability of scores and the validity of their recommended interpretations are available, the information should be included in the test's documentation. When relevant for test interpretation, test documents ordinarily should include item level information, cut scores and configural rules, information about raw scores and derived scores, normative data, the standard errors of measurement, and a description of the procedures used to equate multiple forms.

Standard 6.6

When a test relates to a course of training or study, a curriculum, a textbook, or packaged

instruction, the documentation should include an identification and description of the course or instructional materials and should indicate the year in which these materials were prepared.

Standard 6.7

Test documents should specify qualifications that are required to administer a test and to interpret the test scores accurately.

Comment: Statements of user qualifications need to specify the training, certification, competencies, or experience needed to have access to a test.

Standard 6.8

If a test is designed to be scored or interpreted by test takers, the publisher and test developer should provide evidence that the test can be accurately scored or interpreted by the test takers. Tests that are designed to be scored and interpreted by the test taker should be accompanied by interpretive materials that assist the individual in understanding the test scores and that are written in language that the test taker can understand.

Standard 6.9

Test documents should cite a representative set of the available studies pertaining to general and specific uses of the test.

Comment: Summaries of cited studies—excluding published works, dissertations, or proprietary documents—should be made available on request to test users and researchers by the publisher.

Standard 6.10

Interpretive materials for tests, that include case studies, should provide examples illustrating the diversity of prospective test takers.

Comment: For some instruments, the presentation of case studies that are intended to

STANDARDS

assist the user in the interpretation of the test scores and profiles also will be appropriate for inclusion in the test documentation. For example, case studies might cite as appropriate examples of women and men of different ages; individuals differing in sexual orientation; persons representing various ethnic, cultural, or racial groups; and individuals with special needs. The inclusion of examples illustrating the diversity of prospective test takers is not intended to promote interpretation of test scores in a manner inconsistent with legal requirements that may restrict certain practices in some contexts, such as employee selection.

Standard 6.11

If a test is designed so that more than one method can be used for administration or for recording responses—such as marking responses in a test booklet, on a separate answer sheet, or on a computer keyboard—then the manual should clearly document the extent to which scores arising from these methods are interchangeable. If the results are not interchangeable, this fact should be reported, and guidance should be given for the interpretation of scores obtained under the various conditions or methods of administration.

Standard 6.12

Publishers and scoring services that offer computer-generated interpretations of test scores should provide a summary of the evidence supporting the interpretations given.

Comment: The test user should be informed of any cut scores or configural rules necessary for understanding computer-generated score interpretations. A description of both the samples used to derive cut scores or configural rules and the methods used to derive the cut scores should be provided. When proprietary interests result in the withholding of cut scores or configural rules, the owners of the intellectual

property are responsible for documenting evidence in support of the validity of computer-generated score interpretations. Such evidence might be provided, for example, by reporting the finding of an independent review of the algorithms by qualified professionals.

Standard 6.13

When substantial changes are made to a test, the test's documentation should be amended, supplemented, or revised to keep information for users current and to provide useful additional information or cautions.

Standard 6.14

Every test form and supporting document should carry a copyright date or publication date.

Comment: During the operational life of a test, new or revised test forms may be published, and manuals and other materials may be added or revised. Users and potential users are entitled to know the publication dates of various documents that include test norms. Communication among researchers is hampered when the particular test documents used in experimental studies are ambiguously referenced in research reports.

Standard 6.15

Test developers, publishers, and distributors should provide general information for test users and researchers who may be required to determine the appropriateness of an intended test use in a specific context. When a particular test use cannot be justified, the response to an inquiry from a prospective test user should indicate this fact clearly. General information also should be provided for test takers and legal guardians who must provide consent prior to a test's administration.

PART II

**Fairness
in Testing**

A large, stylized Roman numeral 'II' is positioned on the right side of the page. It is composed of a dense grid of small, light-colored characters, creating a textured, halftone-like appearance. The numeral is centered vertically relative to the 'PART II' and 'Fairness in Testing' text.

7. FAIRNESS IN TESTING AND TEST USE

Background

This chapter addresses overriding issues of fairness in testing. It is intended both to emphasize the importance of fairness in all aspects of testing and assessment and to serve as a context for the technical standards. Later chapters address in greater detail some fairness issues involving the responsibilities of test users, the rights and responsibilities of test takers, the testing of individuals of diverse linguistic backgrounds, and the testing of those with disabilities. Chapters 12 through 15 also address some fairness issues specific to psychological, educational, employment and credentialing, and program evaluation applications of testing and assessment.

Concern for fairness in testing is pervasive, and the treatment accorded the topic here cannot do justice to the complex issues involved. A full consideration of fairness would explore the many functions of testing in relation to its many goals, including the broad goal of achieving equality of opportunity in our society. It would consider the technical properties of tests, the ways test results are reported, and the factors that are validly or erroneously thought to account for patterns of test performance for groups and individuals. A comprehensive analysis would also examine the regulations, statutes, and case law that govern test use and the remedies for harmful practices. The *Standards* cannot hope to deal adequately with all these broad issues, some of which have occasioned sharp disagreement among specialists and other thoughtful observers. Rather, the focus of the *Standards* is on those aspects of tests, testing, and test use that are the customary responsibilities of those who make, use, and interpret tests, and that are characterized by some measure of professional and technical consensus.

Absolute fairness to every examinee is impossible to attain, if for no other reasons than the facts that tests have imperfect reliability and that validity in any particular context is a matter of degree. But neither is any alternative selection or evaluation mechanism perfectly fair. Properly designed and used, tests can and do further societal goals of fairness and equality of opportunity. Serious technical deficiencies in test design, use, or interpretation should, of course, be addressed, but the fairness of testing in any given context must be judged relative to that of feasible test and nontest alternatives. It is general practice that large-scale tests are subjected to careful review and empirical checks to minimize bias. The amount of explicit attention to fairness in the design of well-made tests compares favorably to that of many alternative selection or evaluation methods.

It is also crucial to bear in mind that test settings are interpersonal. The interaction of examiner with examinee should be professional, courteous, caring, and respectful. In most testing situations, the roles of examiner and examinee are sharply unequal in status. A professional's inferences and reports from test findings may markedly impact the life of the person who is examined. Attention to these aspects of test use and interpretation is no less important than more technical concerns.

As is emphasized in professional education and training, users of tests should be alert to the possibility that human issues involving examiner and examinee may sometimes affect test fairness. Attention to interpersonal issues is always important, perhaps especially so when examinees have a disability or differ from the examiner in ethnic, racial, or religious background; in gender or sexual orientation; in socioeconomic status; in age; or in other respects that may affect the examinee-examiner interaction.

Varying Views of Fairness

The term *fairness* is used in many different ways and has no single technical meaning. It is possible that two individuals may endorse fairness in testing as a desirable social goal, yet reach quite different conclusions about the fairness of a given testing program. Outlined below are four principal ways in which the term fairness is used. It should be noted, however, that many additional interpretations may be found in the technical and popular literature.

The first two characterizations presented here relate fairness to absence of bias and to equitable treatment of all examinees in the testing process. There is broad consensus that tests should be free from bias (as defined below) and that all examinees should be treated fairly in the testing process itself (e.g., afforded the same or comparable procedures in testing, test scoring, and use of scores). The third characterization of test fairness addresses the equality of testing outcomes for examinee subgroups defined by race, ethnicity, gender, disability, or other characteristics. The idea that fairness requires equality in overall passing rates for different groups has been almost entirely repudiated in the professional testing literature. A more widely accepted view would hold that examinees of equal standing with respect to the construct the test is intended to measure should on average earn the same test score, irrespective of group membership. Unfortunately, because examinees' levels of the construct are measured imperfectly, this requirement is rarely amenable to direct examination. The fourth definition of fairness relates to equity in opportunity to learn the material covered in an achievement test. There would be general agreement that adequate opportunity to learn is clearly relevant to some uses and interpretations of achievement tests and clearly irrelevant to others, although disagreement might arise as to the relevance of opportunity to learn to test fairness in some specific situations.

FAIRNESS AS LACK OF BIAS

Bias is used here as a technical term. It is said to arise when deficiencies in a test itself or the manner in which it is used result in different meanings for scores earned by members of different identifiable subgroups. When evidence of such deficiencies is found at the level of item response patterns for members of different groups, the terms *item bias* or *differential item functioning* (DIF) are often used. When evidence is found by comparing the patterns of association for different groups between test scores and other variables, the term *predictive bias* may be used. The concept of bias and techniques for its detection are discussed below and are also discussed in other chapters of the *Standards*. There is general consensus that consideration of bias is critical to sound testing practice.

FAIRNESS AS EQUITABLE TREATMENT IN THE TESTING PROCESS

There is consensus that just treatment throughout the testing process is a necessary condition for test fairness. There is also consensus that fair treatment of all examinees requires consideration not only of a test itself, but also the context and purpose of testing and the manner in which test scores are used. A well-designed test is not intrinsically fair or unfair, but the use of the test in a particular circumstance or with particular examinees may be fair or unfair. Unfairness can have individual and collective consequences.

Regardless of the purpose of testing, fairness requires that all examinees be given a comparable opportunity to demonstrate their standing on the construct(s) the test is intended to measure. Just treatment also includes such factors as appropriate testing conditions and equal opportunity to become familiar with the test format, practice materials, and so forth. In situations where individual or group test results are reported, just treatment also implies that such reporting should be accurate and fully informative.

Fairness also requires that all examinees be afforded appropriate testing conditions. Careful standardization of tests and administration conditions generally helps to assure that examinees have comparable opportunity to demonstrate the abilities or attributes to be measured. In some cases, however, aspects of the testing process that pose no particular challenge for most examinees may prevent specific groups or individuals from accurately demonstrating their standing with respect to the construct of interest (e.g., due to disability or language background). In some instances, greater comparability may sometimes be attained if standardized procedures are modified. There are contexts in which some such modifications are forbidden by law and other contexts in which some such modifications are required by law. In all cases, standardized procedures should be followed for all examinees unless explicit, documented accommodations have been made.

Ideally, examinees would also be afforded equal opportunity to prepare for a test. Examinees should in any case be afforded equal access to materials provided by the testing organization and sponsor which describe the test content and purpose and offer specific familiarization and preparation for test taking. In addition to assuring equity in access to accepted resources for test preparation, this principle covers test security for nondisclosed tests. If some examinees were to have prior access to the contents of a secure test, for example, basing decisions upon the relative performance of different examinees would be unfair to others who did not have such access. On tests that have important individual consequences, all examinees should have a meaningful opportunity to provide input to relevant decision makers if procedural irregularities in testing are alleged, if the validity of the individual's score is challenged or may not be reported, or if similar special circumstances arise.

Finally, the conception of fairness as equitable treatment in the testing process extends to the reporting of individual and group test results. Individual test score information is entitled to confidential treatment in most circumstances. Confidentiality should be respected; scores should be disclosed only as appropriate. When test scores are reported, either for groups or individuals, score reports should be accurate and informative. It may be especially important when reporting results to nonprofessional audiences to use appropriate language and wording and to try to design reports to reduce the likelihood of inappropriate interpretations. When group achievement differences are reported, for example, including additional information to help the intended audience understand confounding factors such as unequal educational opportunity may help to reduce misinterpretation of test results and increase the likelihood that tests will be used wisely.

FAIRNESS AS EQUALITY IN OUTCOMES OF TESTING

The idea that fairness requires overall passing rates to be comparable across groups is not generally accepted in the professional literature. Most testing professionals would probably agree that while group differences in testing outcomes should in many cases trigger heightened scrutiny for possible sources of test bias, outcome differences across groups do not in themselves indicate that a testing application is biased or unfair. It might be argued that when tests are used for selection, persons who all would perform equally well on the criterion measure if selected should have an equal chance of being chosen regardless of group membership. Unfortunately, there is rarely any direct procedure for determining whether this ideal has been met. Moreover, if score distributions differ from one group to another, it is generally impossible to satisfy this ideal using any test that has a less than perfect correlation with the criterion measure.

Many testing professionals would agree that if a test is free of bias and examinees have received fair treatment in the testing process, then the conditions of fairness have been met. That is, given evidence of the validity of intended test uses and interpretations, including evidence of lack of bias and attention to issues of fair treatment, fairness has been established regardless of group-level outcomes. This view need not imply that unequal testing outcomes should be ignored altogether. They may be important in generating new hypotheses about bias and fair treatment. But in this view, unequal outcomes at the group level have no direct bearing on questions of test fairness. There may be legal requirements to investigate certain differences in outcomes of testing among subgroups. Those requirements further may provide that, other things being equal, a testing alternative that minimizes outcome differences across relevant subgroups should be used. The standards in this chapter are intended to be applied in a manner consistent with legal and regulatory standards.

FAIRNESS AS OPPORTUNITY TO LEARN

This final conception of fairness arises in connection with educational achievement testing. In many contexts, achievement tests are intended to assess what a test taker knows or can do as a result of formal instruction. When some test takers have not had the opportunity to learn the subject matter covered by the test content, they are likely to get low scores. The test score may accurately reflect what the test taker knows and can do, but low scores may have resulted in part from not having had the opportunity to learn the material tested as well as from having had the opportunity and having failed to learn. When test takers have not had the opportunity to learn the material tested, the policy of using their test scores as a basis for withholding a high school diploma, for example, is viewed as unfair. This issue is further discussed in chapter 13, on educational testing.

At least three important difficulties arise with this conception of fairness. First, the definition of *opportunity to learn* is difficult in practice, especially at the level of individuals. Opportunity is a matter of degree. Moreover, the measurement of some important learning outcomes may require students to work with material they have not seen before. Second, even if it is possible to document the topics included in the curriculum for a group of students, specific content coverage for any one student may be impossible to determine. Finally, there is a well-founded desire to assure that credentials attest to certain proficiencies or capabilities. Granting a diploma to a low-scoring examinee on the grounds that the student had insufficient opportunity to learn the material tested means certifying someone who has not attained the degree of proficiency the diploma is intended to signify.

It should be noted that opportunity to learn ordinarily plays no role in determining the fairness of tests used for employment and credentialing, which are covered in chapter 14, nor of admissions testing. In those circumstances, it is deemed fair that the test should cover the full range of requisite knowledge and skills. However, there are situations in which the agency that determines the contents of a test used for employment or credentialing also sets the curriculum that must be followed in preparing to take the test. In such cases, it is the responsibility of that agency to assure that what is to be tested is fully included in the specification of what is to be taught.

Bias Associated With Test Content and Response Processes

The term *bias* in tests and testing refers to construct-irrelevant components that result in systematically lower or higher scores for identifiable groups of examinees. Such construct-irrelevant score components may be introduced due to inappropriate sampling of

test content or lack of clarity in test instructions. They may also arise if scoring criteria fail to credit fully some correct problem approaches or solutions that are more typical of one group than another. Evidence of these potential sources of bias may be sought in the content of the tests, in comparisons of the internal structure of test responses for different groups, and in comparisons of the relationships of test scores to other measures, although none of these types of evidence is unequivocal.

CONTENT-RELATED SOURCES OF TEST BIAS

Bias due to inappropriate selection of test content may sometimes be detected by inspection of the test itself. In some testing contexts, it is common for test developers to engage an independent panel of diverse experts to review test content for language that might be interpreted differently by members of different groups and for material that might be offensive or emotionally disturbing to some test takers. For performance assessments, panels are often engaged to review the scoring rubric as well. A test intended to measure verbal analogical reasoning, for example, should include words in general use, not words and expressions associated with particular disciplines, occupations, ethnic groups, or locations. Where material likely to be differentially interesting or relevant to some examinees is included, it may be balanced by material that may be of particular interest to the remaining examinees.

In educational achievement testing, alignment with curriculum may bear on questions of content-related test bias. One may ask how well a test represents some content domain and also whether that domain is appropriate given intended score interpretations. A test of 19th-century United States history might give considerable emphasis to the War of 1812, the Mexican War, the Civil War, and the Spanish American War. If some state's curriculum framework dealt relatively

lightly with these wars, devoting more attention instead, say, to social and industrial developments, then that state's test takers might be relatively disadvantaged.

Bias may also result from a lack of clarity in test instructions or from scoring rubrics that credit responses more typical of one group than another. For example, cognitive ability tests often require test takers to classify objects according to an unspecified rule. If a given task credits classification on the basis of the stimulus objects' functions, but an identifiable subgroup of examinees tends to classify the objects on the basis of their physical appearance, faulty test interpretations are likely. Similarly, if the scoring rubric for a constructed response item reserves the highest score level for those examinees who in fact provide more information or elaboration than was actually requested, then less test-wise examinees who simply follow instructions will earn lower scores. In this case, testwiseness becomes a construct-irrelevant component of test scores.

Judgmental methods for the review of tests and test items are often supplemented by statistical procedures for identifying items on tests that function differently across identifiable subgroups of examinees. Differential item functioning (DIF) is said to exist when examinees of equal ability differ on average, according to their group membership, in their responses to a particular item. If examinees from each group are divided into subgroups according to the tested ability and subgroups at the same ability level have unequal probabilities of answering a given item correctly, then there is evidence that that item may not be functioning as intended. It may be measuring something different from the remainder of the test or it may be measuring with different levels of precision for different subgroups of examinees. Such an item may offer a valid measurement of some narrow element of the intended construct, or it may tap some construct-irrelevant component that advantages

or disadvantages members of one group. Although DIF procedures may hold some promise for improving test quality, there has been little progress in identifying the causes or substantive themes that characterize items exhibiting DIF. That is, once items on a test have been statistically identified as functioning differently from one examinee group to another, it has been difficult to specify the reasons for the differential performance or to identify a common deficiency among the identified items.

RESPONSE-RELATED SOURCES OF TEST BIAS

In some cases, construct-irrelevant score components may arise because test items elicit varieties of responses other than those intended or can be solved in ways that were not intended. For example, clients responding to a diagnostic inventory may attempt to provide the answers they think the test administrator expects as opposed to the answers that best describe themselves. To the extent that such response acquiescence is more typical of some groups than others, bias may result. Bias may also be associated with test response formats that pose particular difficulties for one group or another. For example, test performance may rely on some capability (e.g., English language proficiency or fine-motor coordination) that is irrelevant to the intent of the measurement but nonetheless poses impediments for some examinees. A test of quantitative reasoning that makes inappropriately heavy demands on verbal ability would probably be biased against examinees whose first language is other than that of the test.

In addition to content reviews and DIF analyses, evidence of bias related to response processes may be provided by comparisons of the internal structure of the test responses for different groups of examinees. If an analysis of the factors or dimensions underlying test performance reveals different internal structures for different groups, it may be that different constructs are being measured or it

may simply be that groups differ in their variability with respect to the same underlying dimensions. When there is evidence that tests, including personality tests, measure different constructs in different gender, racial, or cultural groups, it is important to determine that the internal structure of the test supports inferences made for clients from these distinct subgroups of the client population. In situations where internal test structure varies markedly across ethnically diverse cultures, it may be inappropriate to make direct comparisons of scores of members of these different cultural groups.

Bias may also be indicated by patterns of association between test scores and other variables. Perhaps the most familiar form such evidence may take is a difference across groups in the regression equations relating selection test performance to criterion performance. This case is discussed at greater length in the following section. However, evidence of bias based on relations to other variables may also take many other forms. The relationship between two tests of the same cognitive ability might be found to differ from one group to another, for example. Such a difference might indicate bias in one or both tests. As another instance, a higher than expected association between reading and mathematics achievement test scores among students who might well have limited English proficiency could trigger an investigation to determine whether language proficiency was influencing some examinees' mathematics scores. Patterns of score averages or other distributional summaries might also point to potential sources of test bias. If males outperformed females on one measure of academic performance and, in the same population, females outperformed males on another, it would follow that the two measures could not both be linearly related to the identical underlying construct. Note, however, that if the tested populations differed, if the content domains sampled differed, or if

the constructs tested otherwise differed due to varying motivational contexts or other effects, two reliable tests, each valid for its intended purpose, might show such a pattern. Association need not imply any direct or causal linkage, and alternative explanations for patterns of association should usually be considered. In some cases, a test-criterion correlation may arise because the test and criterion both depend on the same construct-irrelevant ability. If identifiable subgroups differ with respect to that extraneous ability, then bias may result.

Fairness in Selection and Prediction

When tests are used for selection and prediction, evidence of bias or lack of bias is generally sought in the relationships between test and criterion scores for the respective groups. Under one broadly accepted definition, no bias exists if the regression equations relating the test and the criterion are indistinguishable for the groups in question. (Some formulations may hold that not only regression slopes and intercepts but also standard errors of estimate must be equal.) If test-criterion relationships differ, different decision rules may be followed depending on the group to which the person belongs.

If fitting a common prediction equation for all groups combined suggests that the criterion performance of persons in any one group is systematically overpredicted or underpredicted, and if bias in the criterion measure has been set aside as a possible explanation, one possibility is to generate a separate prediction formula for each group. Another possibility is to seek predictor variables that may be used in lieu of or in addition to the initial predictor score to reduce differential prediction without reducing overall predictive accuracy. If separate regression equations are employed, the effect of their use on the distribution of predicted criterion

scores for the different groups should be examined. Note that in the United States, the use of different selection rules for identifiable subgroups of examinees is legally proscribed in some contexts. There may, however, be legal requirements to consider alternative selection procedures in some such situations.

There is often tension between the perspective that equates fairness with lack of bias, in the technical sense, and the perspective that focuses on testing outcomes. A test that is valid for its intended purpose might be considered fair if a given test score predicts the same performance level for members of all groups. It might nonetheless be regarded by some as unfair, however, if average test scores differ across groups. This is because a given selection score and criterion threshold will often result in proportionately more false negative decisions in groups with lower mean test scores. In other words, a lower-scoring group will usually have a higher proportion of examinees who are rejected on the basis of their test scores even though they would have performed successfully if they had been selected. This seeming paradox is a statistical consequence of the imperfect correlation between test and criterion. It does not occur because of any other property of the test and has no direct relationship to group demographics. It is a purely statistical phenomenon that occurs as a function of lower test scores, regardless of group membership. For example, it usually occurs when the top and bottom test score halves of the majority group are compared. The fairness of a test or another predictor should be evaluated relative to that of nontest alternatives that might be used instead.

GROUP OUTCOME DIFFERENCES DUE TO CHOICE OF PREDICTORS

Success in virtually all real-world endeavors requires multiple skills and abilities, which may interact in complex ways. Testing programs typically address only a

STANDARDS

subset of these. Some skills and abilities are excluded because they are assessed in other components of the selection process (e.g., completion of course work or an interview); others may be excluded because reliable and valid measurement is economically, logistically, or administratively infeasible. Success in college, for example, requires perseverance, motivation, good study habits, and a host of other factors in addition to verbal and quantitative reasoning ability. Even if each of the criteria employed in a selection process is demonstrably valid and appropriate for that purpose, issues of fairness may arise in the choice of which factors are measured. If identifiable groups differ in their average levels of measured versus unmeasured job-relevant characteristics, then fairness becomes a concern at the group level as well as the individual level.

Can Consensus Be Achieved?

It is unlikely that consensus in society at large or within the measurement community is imminent on all matters of fairness in the use of tests. As noted earlier, fairness is defined in a variety of ways and is not exclusively addressed in technical terms; it is subject to different definitions and interpretations in different social and political circumstances. According to one view, the conscientious application of an unbiased test in any given situation is fair, regardless of the consequences for individuals or groups. Others would argue that fairness requires more than satisfying certain technical requirements. It bears repeating that while the *Standards* will provide more specific guidance on matters of technical adequacy, matters of values and public policy are crucial to responsible test use.

Standard 7.1

When credible research reports that test scores differ in meaning across examinee subgroups for the type of test in question, then to the extent feasible, the same forms of validity evidence collected for the examinee population as a whole should also be collected for each relevant subgroup. Subgroups may be found to differ with respect to appropriateness of test content, internal structure of test responses, the relation of test scores to other variables, or the response processes employed by individual examinees. Any such findings should receive due consideration in the interpretation and use of scores as well as in subsequent test revisions.

Comment: Scores differ in meaning across subgroups when the same score produces systematically different inferences about examinees who are members of different subgroups. In those circumstances where credible research reports differences in score meaning for particular subgroups for the type of test in question, this standard calls for separate, parallel analyses of data for members of those subgroups, sample sizes permitting. Relevant examinee subgroups may be defined by race or ethnicity, culture, language, gender, disability, age, socioeconomic status, or other classifications. Not all forms of evidence can be examined separately for members of all such groups. The validity argument may rely on existing research literature, for example, and such literature may not be available for some populations. For some kinds of evidence, some separate subgroup analyses may not be feasible due to the limited number of cases available. Data may sometimes be accumulated so that these analyses can be performed after the test has been in use for a period of time. This standard is not satisfied by assuring that such groups are represented within larger, pooled samples, although this

may also be important. In giving “due consideration in the interpretation and use of scores,” pursuant to this standard, test users should be mindful of legal restrictions that may prohibit or limit within-group scoring and other practices.

Standard 7.2

When credible research reports differences in the effects of construct-irrelevant variance across subgroups of test takers on performance on some part of the test, the test should be used if at all only for those subgroups for which evidence indicates that valid inferences can be drawn from test scores.

Comment: An obvious reason why a test may not measure the same constructs across subgroups is that different components come into play from one subgroup to another. Alternatively, an irrelevant component may have a more significant effect on the performance of examinees in one subgroup than in another. Such intrusive elements are rarely entirely absent for any subgroup but are seldom present to any great extent. The decision whether or not to use a test with any given examinee subgroup necessarily involves a careful analysis of the validity evidence for different subgroups, as called for in Standard 7.1, and the exercise of thoughtful professional judgment regarding the significance of the irrelevant components.

A conclusion that a test is not appropriate for a particular subgroup requires an alternative course of action. This may involve a search for a test that can be used for all groups or, in circumstances where it is feasible to use different construct-equivalent tests for different groups, for an alternative test for use in the subgroup for which the intended construct is not well measured by the current test. In some cases multiple tests may be used in combination,

and a composite that permits valid inferences across subgroups may be identified. In some circumstances, such as employment testing, there may be legal or other constraints on the use of different tests for different subgroups.

It is acknowledged that there are occasions where examinees may request or demand to take a version of the test other than that deemed most appropriate by the developer or user. An individual with a disability may decline an alternate form and request the standard form. Acceding to this request, after ensuring that the examinee is fully informed about the test and how it will be used, is not a violation of this standard.

Standard 7.3

When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers in the content domain measured by the test, test developers should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups.

Comment: Differential item functioning exists when examinees of equal ability differ, on average, according to their group membership in their responses to a particular item. In some domains, existing research may indicate that differential item functioning occurs infrequently and does not replicate across samples. In others, research evidence may indicate that differential item functioning occurs reliably at meaningful above-chance levels for some particular groups; it is to such circumstances that the standard applies. Although it may not be possible prior to first release of a test to

STANDARDS

FAIRNESS IN TESTING AND TEST USE / PART II

study the question of differential item functioning for some such groups, continued operational use of a test may afford opportunities to check for differential item functioning.

Standard 7.4

Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain.

Comment: Two issues are involved. The first deals with the inadvertent use of language that, unknown to the test developer, has a different meaning or connotation in one subgroup than in others. Test publishers often conduct sensitivity reviews of all test material to detect and remove sensitive material from the test. The second deals with settings in which sensitive material is essential for validity. For example, history tests may appropriately include material on slavery or Nazis. Tests on subjects from the life sciences may appropriately include material on evolution. A test of understanding of an organization's sexual harassment policy may require employees to evaluate examples of potentially offensive behavior.

Standard 7.5

In testing applications involving individualized interpretations of test scores other than selection, a test taker's score should not be accepted as a reflection of standing on the characteristic being assessed without consideration of alternate explanations for the test taker's performance on that test at that time.

Comment: Many test manuals point out variables that should be considered in interpreting test scores, such as clinically relevant history, school record, vocational status, and test-taker motivation. Influences associated with variables such as socioeconomic status, ethnicity, gender, cultural background, language, or age may also be relevant. In addition, medication, visual impairments, or other disabilities may affect a test taker's performance on, for example, a paper-and-pencil test of mathematics.

Standard 7.6

When empirical studies of differential prediction of a criterion for members of different subgroups are conducted, they should include regression equations (or an appropriate equivalent) computed separately for each group or treatment under consideration or an analysis in which the group or treatment variables are entered as moderator variables.

Comment: Correlation coefficients provide inadequate evidence for or against a differential prediction hypothesis if groups or treatments are found not to be approximately equal with respect to both test and criterion means and variances. Considerations of both regression slopes and intercepts are needed. For example, despite equal correlations across groups, differences in intercepts may be found.

Standard 7.7

In testing applications where the level of linguistic or reading ability is not part of the construct of interest, the linguistic or reading demands of the test should be kept to the minimum necessary for the valid assessment of the intended construct.

Comment: When the intent is to assess ability in mathematics or mechanical comprehension, for example, the test should not contain unusual words or complicated syntactic conventions unrelated to the mathematical or mechanical skill being assessed.

Standard 7.8

When scores are disaggregated and publicly reported for groups identified by characteristics such as gender, ethnicity, age, language proficiency, or disability, cautionary statements should be included whenever credible research reports that test scores may not have comparable meaning across these different groups.

Comment: Comparisons across groups are only meaningful if scores have comparable meaning across groups. The standard is intended as applicable to settings where scores are implicitly or explicitly presented as comparable in score meaning across groups.

Standard 7.9

When tests or assessments are proposed for use as instruments of social, educational, or public policy, the test developers or users proposing the test should fully and accurately inform policymakers of the characteristics of the tests as well as any relevant and credible information that may be available concerning the likely consequences of test use.

Standard 7.10

When the use of a test results in outcomes that affect the life chances or educational opportunities of examinees, evidence of mean test score differences between relevant subgroups of examinees should, where feasible, be examined for subgroups for which credible research reports mean differences for similar tests. Where mean

differences are found, an investigation should be undertaken to determine that such differences are not attributable to a source of construct underrepresentation or construct-irrelevant variance. While initially the responsibility of the test developer, the test user bears responsibility for uses with groups other than those specified by the developer.

Comment: Examples of such test uses include situations in which a test plays a dominant role in a decision to grant or withhold a high school diploma or to promote a student or retain a student in grade. Such an investigation might include a review of the cumulative research literature or local studies, as appropriate. In some domains, such as cognitive ability testing in employment, a substantial relevant research base may preclude the need for local studies. In educational settings, as discussed in chapter 13, potential differences in opportunity to learn may be relevant as a possible source of mean differences.

Standard 7.11

When a construct can be measured in different ways that are approximately equal in their degree of construct representation and freedom from construct-irrelevant variance, evidence of mean score differences across relevant subgroups of examinees should be considered in deciding which test to use.

Comment: Mean score differences, while important, are but one factor influencing the choice between one test and another. Cost, testing time, test security, and logistic issues (e.g., an application where very large numbers of examinees must be screened in a very short time) are among the issues also entering into the professional judgment about test use.

STANDARDS

Standard 7.12

The testing or assessment process should be carried out so that test takers receive comparable and equitable treatment during all phases of the testing or assessment process.

Comment: For example, should a person administering a test or interpreting test results recognize a personal bias for or against an examinee, or for or against any subgroup of which the examinee is a member, the person could take a variety of steps ranging from seeking a review of test interpretations from a colleague to withdrawal from the testing process.

8. THE RIGHTS AND RESPONSIBILITIES OF TEST TAKERS

Background

This chapter addresses fairness issues unique to the interests of the individual test taker. Fair treatment of test takers is not only a matter of equity, but also promotes the validity and reliability of the inferences made from the test performance. The standards presented in this chapter reflect widely accepted principles in the field of measurement. The standards address the responsibilities of test takers with regard to test security, their access to test results, and their rights when irregularities in their testing are claimed. Other issues of fairness are treated in other chapters: general principles in chapter 7; the testing of linguistic minorities in chapter 9; the testing of persons with disabilities in chapter 10. General considerations concerning reports of test results are covered in chapter 5.

Test takers have the right to be assessed with tests that meet current professional standards, including standards of technical quality, fairness, administration, and reporting of results. Fair and equitable treatment of test takers involves providing, in advance of testing, information about the nature of the test, the intended use of test scores, and the confidentiality of the results. Test takers, or their legal representatives when appropriate, need enough information about the test and the intended use of test results to reach a competent decision about participating in testing. In some instances, formal informed consent for testing is required by law or by other standards of professional practice, such as those governing research on human subjects. The greater the consequences to the test taker, the greater the importance of ensuring that the test taker is fully informed about the test and voluntarily consents to participate, except when testing without consent is permitted by law. If a test is optional, the test

taker has the right to know the consequences of taking or not taking the test. The test taker has the right to acceptable opportunities for asking questions or expressing concerns, and may expect timely responses to legitimate questions.

Where consistent with the purposes and nature of the assessment, general information is usually provided about the test's content and purposes. Some programs, in the interests of fairness, provide all test takers with helpful materials, such as study guides, sample questions, or complete sample tests, when such information does not jeopardize the validity of the results from future test administration. Advice may also be provided about test-taking strategies, including time management, and the advisability of omitting an item response, when it is permitted. Information is made known about the availability of special accommodations for those who need them. The policy on retesting may be stated, in case the test taker feels that the present performance does not appropriately reflect his/her best performance.

As participants in the assessment, test takers have responsibilities as well as rights. Their responsibilities include preparing themselves for the test, following the directions of the test administrator, representing themselves honestly on the test, and informing appropriate persons if they believe the test results do not adequately reflect them. In group testing situations, test takers are expected not to interfere with the performance of other test takers.

Test validity rests on the assumption that a test taker has earned fairly a particular score or pass/fail decision. Any form of cheating, or other behavior that reduces the fairness and validity of a test, is irresponsi-

STANDARDS

THE RIGHTS AND RESPONSIBILITIES OF TEST TAKERS / PART II

ble, is unfair to other test takers and may lead to sanctions. It is unfair for a test taker to use aids that are prohibited. It is unfair for a test taker to arrange for someone else to take the test in his/her place. The test taker is obligated to respect the copyrights of the test publisher or sponsor on all test materials. This means that the test taker will not reproduce the items without authorization nor disseminate, in any form, material that is clearly analogous to the reproduction of the items. Test takers, as well as test administrators, have the responsibility not to compromise security by divulging any details of the test items to others nor may they request such details from others. Failure to honor these responsibilities may compromise the validity of test score interpretations for themselves and for others.

Sometimes, testing programs use special scores, statistical indicators, and other indirect information about irregularities in testing to help ensure that the test scores are obtained fairly. Unusual patterns of responses, large changes in test scores upon retesting, speed of responding, and similar indicators may trigger careful scrutiny of certain testing protocols. The details of these procedures are generally kept secure to avoid compromising their use. However, test takers can be made aware that in special circumstances, such as response or test score anomalies, their test responses may get special scrutiny. If evidence of impropriety or fraud so warrants, the test taker's score may be canceled, or other action taken.

Because these *Standards* are directed to test providers, and not to test takers, standards about test-taker responsibilities are phrased in terms of providing information to test takers about their rights and responsibilities. Providing this information is the joint responsibility of the test developer, the test administrator, the test proctor, if any, and the test user and may be apportioned according to particular circumstances.

Standard 8.1

Any information about test content and purposes that is available to any test taker prior to testing should be available to *all* test takers. Important information should be available free of charge and in accessible formats.

Comment: The intent of this standard is equal treatment for all. Important information would include that necessary for testing, such as when and where the test is given, what material should be brought, the purpose of the test, and so forth. More detailed information, such as practice materials, is sometimes offered for a fee. Such offerings should be made to all test takers.

Standard 8.2

Where appropriate, test takers should be provided, in advance, as much information about the test, the testing process, the intended test use, test scoring criteria, testing policy, and confidentiality protection as is consistent with obtaining valid responses.

Comment: Where appropriate, test takers should be informed, possibly by a test bulletin or similar procedure, about test content, including subject area, topics covered, and item formats. They should be informed about the advisability of omitting responses. They should be aware of any imposed time limits, so that they can manage their time appropriately. General advice should be given about test-taking strategy. In computer administrations, they should be told about any provisions for review of items they have previously answered or omitted. Test takers should understand the intended use of test scores and the confidentiality of test results. They should be advised whether they will have access to their results. They should be informed about the policy con-

cerning taking the test again and about the possibility that some test protocols may receive special scrutiny for security reasons. Test takers should be informed about the consequences of misconduct or improper behavior, such as cheating, that could result in their being prohibited from completing the test, receiving test scores, or other sanctions.

Standard 8.3

When the test taker is offered a choice of test format, information about the characteristics of each format should be provided.

Comment: Test takers sometimes have to choose between a paper-and-pencil administration and a computer-administered test, which may be adaptive. Some tests are offered in several different languages. Sometimes an alternative assessment is offered in lieu of the ordinary test. Test takers need to know the characteristics of each alternative so that they can make an informed choice.

Standard 8.4

Informed consent should be obtained from test takers, or their legal representatives when appropriate, before testing is done except (a) when testing without consent is mandated by law or governmental regulation, (b) when testing is conducted as a regular part of school activities, or (c) when consent is clearly implied.

Comment: Informed consent implies that the test takers or representatives are made aware, in language that they can understand, of the reasons for testing, the type of tests to be used, the intended use, and the range of material consequences of the intended use. If written, video, or audio records are made of the testing session, or other records are kept, test takers

are entitled to know what testing information will be released and to whom. Consent is not required when testing is legally mandated, such as a court-ordered psychological assessment, but there may be legal requirements for providing information. When testing is required for employment or for educational admissions, applicants, by applying, have implicitly given consent to the testing. Nevertheless, test takers and/or their legal representatives should be given appropriate information about a test when it is in their interest to be informed. Young test takers should receive an explanation of the reasons for testing. Even a child as young as two or three, as well as older test takers of limited cognitive ability, can understand a simple explanation as to why they are being tested (such as, "I'm going to ask you to try to do some things so that I can see what you know how to do and what things you could use some more help with").

Standard 8.5

Test results identified by the names of individual test takers, or by other personally identifying information, should be released only to persons with a legitimate, professional interest in the test taker or who are covered by the informed consent of the test taker or a legal representative, unless otherwise required by law.

Comment: Scores of individuals identified by name, or by some other means by which a person can be readily identified, such as social security number, should be kept confidential. In some situations, information may be provided on a confidential basis to other practitioners with a legitimate interest in the particular case, consistent with legal and ethical considerations. Information may be provided to researchers if a test taker's anonymity is maintained and the

STANDARDS

THE RIGHTS AND RESPONSIBILITIES OF TEST TAKERS / PART II

intended use is consistent with accepted research practice and is not inconsistent with the conditions of the test taker's informed consent.

Standard 8.6

Test data maintained in data files should be adequately protected from improper disclosure. Use of facsimile transmission, computer networks, data banks, and other electronic data processing or transmittal systems should be restricted to situations in which confidentiality can be reasonably assured.

Comment: When facsimile or computer communication is used to transmit a test protocol to another site for scoring, or if scores are similarly transmitted, special provisions should be made to keep the information confidential. See Standard 5.13.

Standard 8.7

Test takers should be made aware that having someone else take the test for them, disclosing confidential test material, or any other form of cheating is inappropriate and that such behavior may result in sanctions.

Comment: Although the standards cannot regulate the behavior of test takers, test takers should be made aware of their personal and legal responsibilities. Arranging for someone else to impersonate the nominal test taker constitutes fraud. Disclosure of confidential testing material for the purpose of giving other test takers pre-knowledge is unfair and may constitute copyright infringement. In licensure and certification tests, such actions may compromise public health and safety. The validity of test score interpretations is compromised by inappropriate test disclosure.

Standard 8.8

When score reporting includes assigning individuals to categories, the categories should be chosen carefully and described precisely. The least stigmatizing labels, consistent with accurate representation, should always be assigned.

Comment: When labels are associated with test results, care should be taken to be precise in the meanings associated with the labels and to avoid unnecessarily stigmatizing consequences associated with a label. For example, in an assessment designed to aid in determining whether an individual is competent to stand trial, the label "incompetent" is appropriate for individuals who perform poorly on the assessment. However, in a test of basic literacy skills, it is more appropriate to use a label such as "not proficient" rather than "incompetent," because the latter term has a more global and derogatory meaning.

Standard 8.9

When test scores are used to make decisions about a test taker or to make recommendations to a test taker or a third party, the test taker or the legal representative is entitled to obtain a copy of any report of test scores or test interpretation, unless that right has been waived or is prohibited by law or court order.

Comment: In some cases a test taker may be adequately informed when the test report is given to an appropriate third party (treating psychologist or psychiatrist) who can interpret the findings to the test taker. In professional applications of individualized testing, when the test taker is given a copy of the test report, the examiner or a knowledgeable third party should be available to interpret it, even if it is clearly written, as the test

taker may misunderstand or raise questions not specifically answered in the report. In employment testing situations, where test results are used solely for the purpose of aiding selection decisions, waivers of access are often a condition of employment, although access to test information may often be appropriately required in other circumstances.

Standard 8.10

In educational testing programs and in licensing and certification applications, when an individual score report is expected to be delayed beyond a brief investigative period, because of possible irregularities such as suspected misconduct, the test taker should be notified, the reason given, and reasonable efforts made to expedite review and to protect the interests of the test taker. The test taker should be notified of the disposition, when the investigation is closed.

Standard 8.11

In educational testing programs and in licensing and certification applications, when it is deemed necessary to cancel or withhold a test taker's score because of possible testing irregularities, including suspected misconduct, the type of evidence and procedures to be used to investigate the irregularity should be explained to all test takers whose scores are directly affected by the decision. Test takers should be given a timely opportunity to provide evidence that the score should not be canceled or withheld. Evidence considered in deciding upon the final action should be made available to the test taker on request.

Comment: Any form of cheating or behavior that reduces the validity and fairness of test results should be investigated promptly, and

appropriate action taken. Withholding or canceling a test score may arise because of suspected misconduct by the test taker, or because of some anomaly involving others, such as theft, or administrative mishap. An avenue of appeal should be available and made known to candidates whose scores may be amended or withheld. Some testing organizations offer the option of a prompt and free retest or arbitration of disputes.

Standard 8.12

In educational testing programs and in licensing and certification applications, when testing irregularities are suspected, reasonably available information bearing directly on the assessment should be considered, consistent with the need to protect the privacy of test takers.

Comment: Unless allegations of misconduct are made by associates of the test taker, the information to be collected would ordinarily be limited to that obtainable without invading the privacy of the test taker or his/her associates.

Standard 8.13

In educational testing programs and in licensing and certification applications, test takers are entitled to fair consideration and reasonable process, as appropriate to the particular circumstances, in resolving disputes about testing. Test takers are entitled to be informed of any available means of recourse.

Comment: When a test taker's score may be questioned and may be invalidated, or when a test taker seeks a review or revision of his/her score or some other aspect of the testing, scoring, or reporting process, the test taker is entitled to some orderly process for effective input into or review of the

STANDARDS

THE RIGHTS AND RESPONSIBILITIES OF TEST TAKERS / PART II

decision making of the test administrator or test user. Depending upon the magnitude of the consequences associated with the test, this can range from an internal review of all relevant data by a test administrator, to an informal conversation with an examinee, to a full administrative hearing. The greater the consequences, the greater the extent of procedural protections that should be made available. Test takers should also be made aware of procedures for recourse, fees, expected time for resolution, and any possible consequences for the test taker. Some testing programs advise that the test taker may be represented by an attorney, although possibly at the test taker's expense.

UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA

AMERICAN EDUCATIONAL RESEARCH)	
ASSOCIATION, INC., AMERICAN)	
PSYCHOLOGICAL ASSOCIATION, INC.,)	
and NATIONAL COUNCIL ON)	
MEASUREMENT IN EDUCATION, INC.,)	Civil Action No. 1:14-cv-00857-TSC-DAR
)	
Plaintiffs,)	DECLARATION OF KURT F.
)	GEISINGER IN SUPPORT OF
v.)	PLAINTIFFS’ MOTION FOR
)	SUMMARY JUDGMENT AND ENTRY
PUBLIC.RESOURCE.ORG, INC.,)	OF A PERMANENT INJUNCTION
)	
Defendant.)	
)	

I, KURT F. GEISINGER, declare:

1. I am currently Director of the Buros Center on Testing and W. C. Meierhenry Distinguished University Professor at the University of Nebraska-Lincoln. I submit this Declaration in support of the motion of the American Educational Research Association, Inc. (“AERA”), the American Psychological Association, Inc. (“APA”), and the National Council on Measurement in Education, Inc. (“NCME”) (collectively, “Plaintiffs” or “Sponsoring Organizations”) for summary judgment and the entry of a permanent injunction.

2. My curriculum vitae is attached to this Declaration as Exhibit 1.

3. I received my doctoral degree in Educational Psychology in 1977 from the Pennsylvania State University, after previously receiving my masters’ degree in Psychology at the University of Georgia and my bachelor’s degree from Davidson College (with honors). I also studied German, Psychology and other topics as an undergraduate at the Phillips Universität in Marburg, Germany and at Harvard University when I attended the Institute for Educational Management in 1995.

4. From 2001 to 2006, I served as the Vice President of Academic Affairs and Professor of Psychology at the University of St. Thomas in Houston, Texas, where I was responsible for four academic schools, approximately 200 faculty members, and over 4,000 students. From 1997 to 2001, I served as Academic Vice President and Professor of Psychology at Le Moyne College. From 1992 to 1997 I served as Dean of the College of Arts and Sciences and Professor of Psychology at the State University of New York at Oswego. And, from 1977-1992 I served as a Professor of Psychology at Fordham University in New York City, where I served as department chair for the Department of Psychology and director of the Doctoral program in Psychometrics.

5. Over the past forty years, I have researched, studied, and taught psychometrics (psychometrics is the quantitative study of tests and measures in terms of the value, usefulness, and interpretation of the results of such measures). I also am a fellow, diplomate, and member of numerous professional societies involving educational and psychological testing, such as the APA (fellow), the American Association for Assessment Psychology (diplomate), the AERA (fellow), and the NCME, as well as other professional associations. I have represented the APA by serving on and chairing the Joint Committee on Testing Practices (which is separate from the APA committee responsible for the 1999 *Standards for Educational and Psychological Testing*) and have served on the APA's Committee on Psychological Testing and Assessment. In 2010, I was elected to serve two terms (2006-2008 and 2009-2011) as the representative on the Council of Representatives for the APA's Division of Evaluation, Measurement and Statistics. My second term was cut short by one year when I was elected to serve as a member-at-large on the APA's Board of Directors in 2010, a position I held for a three-year term (2011-2013).

6. I have authored numerous publications about psychological and educational testing. I have worked at the Educational Testing Service (“ETS”), chaired its Technical Advisory Committee for the Graduate Record Examination (“GRE”), served on the Board of Directors for the GRE (a Board that I also chaired), and have been a member of the College Board, (formerly known as the College Entrance Examination Board) for which I served (2000-2002) on its SAT Committee. I recently concluded a four-year term (2011-2014) on the Advisory Research Committee for the College Board, serving the last two years as its chair. I currently serve on the Technical Advisory Committee for the Educational Records Bureau¹ and on Saudi Arabia’s International Advisory Board for its National Center for Assessment and Evaluation.

7. In 2010, I was elected to the Council (i.e., Board of Directors) for the International Test Commission—the primary international testing body. In 2012, I was also elected as its Treasurer and to serve on its Executive Council. I am the only American who serves on its Executive Council.

8. I was asked to review and share my comments’ chapters of the 1999 *Standards for Educational and Psychological Testing*, published jointly by the AERA, the APA, and the NCME (the “1999 Standards”). The Standards² embody the professionally accepted practices for testing and measurement. One of the chapters I reviewed was based upon the testing of individuals with disabilities, an area in which I have engaged in research and have served as an expert witness in federal courts as well as state courts in New York, New Jersey, and California.

¹ The Educational Record Bureau specializes in the development and use of tests and testing products for private and independent educational institutions at the p-12 levels.

² I use the term “Standards” to refer to the *Standards for Educational and Psychological Testing* as a whole, not a specific version of the Standards, i.e. 1999 or 2014

The other chapter related to the rights and responsibilities of test takers. *See* Exh. 1. I note that the Standards were revised again in 2014.

9. In addition to my 130 plus journal articles and book chapters, I have written, edited, or co-edited approximately 15 books and monographs. The vast majority of these publications deal with testing and measurement issues. For example, I have edited two books on the psychological testing of Hispanics and another I co-edited related to fairness in testing. I have also co-edited several books of reviews of published tests and measures. I was also Editor-in-Chief for the three-volume *Handbook of Testing and Assessment in Psychology* (published by the APA in 2013) and have been editor of the journal *Applied Measurement in Education* for the past 9 plus years. Taylor & Francis, in conjunction with the Buros Center for Testing publishes this journal.

10. I also co-chaired a sub-committee of the APA's Joint Committee on Testing Practices and the overall committee itself that developed a document on the rights and responsibilities of test takers (1993-2001). This document has been endorsed by a number of professional associations related to proper test use, including the APA, the National Association of School Psychologists, the American Counseling Association, and the NCME. While chairing the Joint Committee on Testing Practices, the committee developed a book entitled *Assessing Individuals with Disabilities*, in which I wrote a chapter. I also served on a task force charged to illuminate issues related to the testing of individuals with disabilities as well as ethnic minorities. The task force wrote and edited a book entitled *Test Interpretation and Diversity: Achieving Equity in Assessment*, which was published by the APA's publication unit in 1997. I had three chapters in that volume.

11. I additionally served on an APA task force (2007-2010) that considered the assessment and intervention of individuals with disabilities. The results of our work, Guidelines for the “Assessment of and Intervention with Individuals with Disabilities,” was published in the *American Psychologist*, the premier publication of the APA (Geisinger et al., 2012) and endorsed as the policy of the APA by its governance. A reference for the *American Psychologist* article may be found on my curriculum vitae, which is attached as Exhibit 1.

12. In the past two years (2014-2015), I have served on two task forces related to the use of measures in clinical psychology. One of these has written a policy, recently accepted by the APA’s Board of Directors, that differentiates the use of tests and other measures, for screening and assessment, two highly related types of testing, but which differ in specificity and focus. Tests are usually standardized measures that are given to a number of people for a specific purpose. A bar examination would be an example of a test. Measures are other variables yielding typically quantitative values that are used to evaluate a person and include tests. A bathroom scale results in a measure (weight), but would not normally be considered as a test.

13. During 2013-2014, I served on a committee of the Institute of Medicine (a component of the National Academy of Sciences) that evaluated the use of psychological and clinical neuropsychological measures by the Social Security Administration in determining disability status. The final report, entitled *Psychological Testing in the Service of Disability Determination*, has been published by the National Academy of Sciences and is also available from the Institute of Medicine’s website.

14. For approximately four years (2008-2012), I jointly represented three professional associations (the AERA, the APA, and the NCME) in developing the International Standards

Organization's ("ISO") first standard on psychological testing. The results of the work of the committee that engaged in this activity was ISO Standard 10677. The standard is divided into two parts. The first establishes requirements and guidance for a client working with a service provider to carry out the assessment of an individual, a group, or an organization for work-related purposes. ISO 10667-1:2011 enables the client to base its decisions on sound assessment results. ISO 10667-1:2011 also specifies the responsibilities of a service provider in terms of the assessment methods and procedures that can be carried out for various work-related purposes made by or affecting individuals, groups or organizations.

15. I also built or helped to build a number of testing measures. Specifically, I served as the primary consultant on a number of civil service examinations given in New York City for police officers, sergeants, lieutenants, and captains, fire fighters, fire lieutenants, fire captains, sanitation supervisors, and a variety of other civil service occupations over a period of at least a decade ending in 1992. I sometimes defended these measures in court. I also represented the Public Service Alliance of Canada against the Public Service of Canada in two cases related to their national testing efforts and assisted Disability Rights Advocates with regard to several testing disputes concerning individuals with disabilities. *See* Exh. 1.

16. In recent years, my primary efforts have been to assure testing fairness for those with disabilities, language minorities, and ethnic minorities.

17. I first learned about the Standards for Educational and Psychological Testing while I was in my first or second year of graduate school. They are widely discussed in classes on testing and testing practice and treated with great respect. Some graduate programs and courses require that students purchase the Standards as part of their coursework and education. In teaching graduate classes on topics related to testing and associated with the Standards, I often

refer to them, building the thoughts and approaches described in the Standards, as well as specific standards, into my lectures and classes. I expect students to purchase and read the Standards in a number of the classes I taught. When writing chapters and articles on such topics as test validity, test reliability, and test fairness—all topics I have discussed in writing—I frequently refer to the Standards to check my use of language, my interpretations, and to check that I am not omitting a topic of importance relevant to the specific publication. Also, when building tests, such as the Police and Fire Department Civil Service tests I helped construct for the City of New York, or when serving on technical advisory committees for the well-known SAT and GRE committees and boards, I refer to the Standards frequently. Usually, in meetings I attempt to express what I believed to be best practices, and then would “back up” my beliefs with quotes from the specific and relevant standards. Perhaps my greatest use of the Standards has occurred in my legal defense of specific tests or in my critique of particular uses of some tests, both of which I have engaged in during my career as an expert witness.

18. The ultimate advantages of the Standards in my opinion are that they are written and edited by first-rate professionals covering a number of the representative fields in which testing and assessment are primarily employed, they are thoroughly and publicly vetted by other professionals, and they are openly discussed during the revision process at many professional conferences. The resultant document becomes a living document of best practices. That the members of the committee drafting the Standards are generally extremely highly respected professionals in the field of testing and testing practice also provides the Standards great credibility. Given my experience over the last 10 years as Director of the Buros Center for Testing, thought of by many as the Consumer Reports of the testing industry, and my service as the co-editor of the Mental Measurements Yearbooks, where commercially available tests are

evaluated, I can state categorically that the Standards serve as the primary basis for all test evaluations. The other editors of these Yearbooks and I refer to the Standards with great frequency to determine and assure ourselves that the comments made by reviewers are consistent with the Standards and that the reviews themselves are based upon principles supported by, and coherent with, the Standards. The Standards originally were created as principles and guidelines – a set of best practices to improve professional practice in testing and assessment across multiple settings, including education and various areas of psychology. The Standards can and should be used as a recommended course of action in the sound and ethical development and use of tests, and also to evaluate the quality of tests and testing practices. Additionally, an essential component of responsible professional practice is maintaining technical competence. Many professional associations also have developed standards and principles of technical practice in assessment. The Sponsoring Organizations’ Standards have been and still are used for this purpose.

19. The Standards, however, are not simply intended for members of the Sponsoring Organizations: AERA, APA, and NCME. The intended audience of the Standards is broad and cuts across audiences with varying backgrounds and different training. For example, the Standards also are intended to guide test developers, sponsors, publishers, and users by providing criteria for the evaluation of tests, testing practices, and the effects of test use. Test user-oriented standards refer to those standards that help test users decide how to choose certain tests, interpret scores, or make decisions based on test results. Test users include clinical or industrial psychologists, research directors, school psychologists, counselors, employment supervisors, teachers, and various administrators who select or interpret tests for their organizations. There is no mechanism, however, to enforce compliance with the Standards on the part of the test

developer or test user. The Standards, moreover, do not attempt to provide psychometric answers to policy or legal questions. They do not themselves set requirements, but serve to distribute best practices and procedures.

20. The Standards apply broadly to a wide range of standardized instruments and procedures that sample an individual's behavior, including tests, assessments, inventories, scales, and other testing vehicles. The Standards apply equally to standardized multiple-choice tests, performance assessments (including tests comprised of only open-ended essays), and hands-on assessments or simulations. The main exceptions are that the Standards do not apply to unstandardized questionnaires (*e.g.*, unstructured behavioral checklists or observational forms), teacher-made tests, and subjective decision processes (*e.g.*, a teacher's evaluation of students' classroom participation over the course of a semester).

21. The Standards have been used to develop testing guidelines for such activities as college admissions, personnel selection, test translations, test user qualifications, and computer-based testing. The Standards also have been widely cited to address technical, professional, and operational norms for all forms of assessments that are professionally developed and used in a variety of settings. The Standards additionally provide a valuable public service to state and federal governments as they voluntarily choose to use them. For instance, each testing company, when submitting proposals for testing administration, instead of relying on a patchwork of local, or even individual and proprietary, testing design and implementation criteria, may rely instead on the Sponsoring Organizations' Standards to afford the best guidance for testing and assessment practices.

22. The Sponsoring Organizations do not keep any of the revenues generated from the sales of the Standards. Rather, the income from these sales is used by the Sponsoring

Organizations to offset their development and production costs and to generate funds for subsequent revisions. This strategy allows the Sponsoring Organizations to develop up-to-date, high quality Standards that otherwise would not be developed due to the time and effort that goes into producing them.

23. Without the sales revenue from prior Standards versions (because – if Public Resource succeeds in this litigation – this publication will be made freely available online), it is extremely unlikely that future updates to the Standards will be undertaken. This well-informed opinion is made because NCME is too small an organization to financially support periodic updates of the Standards, AERA does not have the budget for it, and an insufficient number of psychometricians are members of APA for it to justify the ongoing expenditures. Charging extra membership fees to fund ongoing updates to the Standards would never happen, because the governing bodies of AERA, APA and NCME would not vote for it. If these Sponsoring Organizations ceased updating the Standards, it is unlikely that other organizations would step in and continue the effort. Moreover, there are no other organizations with the expertise in their memberships to populate such a committee or task force.

24. There simply is no way for Plaintiffs to calculate with any degree of certainty the number of university/college professors, students, testing companies and others who would have purchased Plaintiffs' Standards but for their wholesale posting on Defendant's <https://law.resource.org> website and the Internet Archive <http://archive.org> website.

25. In Fiscal Year (“FY”) 2011 to FY 2012, as compared to FY 2011, the Sponsoring Organizations experienced a 34% drop in sales of the 1999 Standards. In FY 2013, sales of the 1999 Standards remained at their low level from the prior fiscal year (*See* F. Levine Declaration, ¶ 18, Exh. 000). For a publication with the longevity of the 1999 Standards, one otherwise

would expect to see a gradual decline in sales year-over-year; not the precipitous drop in sales experienced by the 1999 Standards in 2012 and 2013 - even considering that updated Standards were published in 2014. It is also clear that this drop did not occur due to the expected publication of the 2014 Standards, because they were actually due to be published more than a year earlier. Thus, one would have expected such a drop to occur perhaps in 2010 or 2011.

26. Past harm from Public Resource's infringing activities includes misuse of Plaintiffs' intellectual property without permission, lost sales that cannot be totally accounted for - due to potentially infinite Internet distribution, especially by psychometrics students, and lack of funding that otherwise would have been available for the update of the Sponsoring Organizations' Standards from the 1999 to the 2014 versions.

27. Should Public Resource's infringement be allowed to continue, the harm to the Sponsoring Organizations, and public at large who rely on the preparation and administration of valid, fair and reliable tests, includes: (i) uncontrolled publication of the 1999 Standards without any notice that those guidelines have been replaced by the 2014 Standards; (ii) future unquantifiable loss of revenue from sales of authorized copies of the 1999 Standards (with proper notice that they are no longer the current version) and the 2014 Standards; and (iii) lack of funding for future revisions of the 2014 Standards and beyond.

28. The harm caused to the public by publication of out-of-date Standards (not labeled as such) will be significant, because the testing and assessment fields are constantly changing, given updates in testing technology and ever-evolving collective thought on the validity, reliability and fairness of tests. Members of the public who would be harmed by discontinued updates of the Standards include psychometrics professors, students and professionals, as well as test developers, administrators and takers.

I DECLARE, under the penalty of perjury, that the foregoing is true and correct.

Dated: December 9, 2015

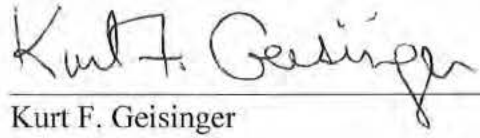

Kurt F. Geisinger

EXHIBIT 1

Updated 12/1/2014

CURRICULUM VITAE

Kurt F. Geisinger, Ph.D.

Current Position: Director, Buros Center for Testing
W. C. Meierhenry Distinguished University Professor
The University of Nebraska-Lincoln

Office Address and Telephone
21 Teachers College Hall
Buros Center for Testing
The University of Nebraska-Lincoln
Lincoln, NE 68588
Telephone: 402/472-3280
FAX: 402/472-6207
E-Mail: kgeisinger2@unl.edu

Home Address and Telephone
6300 Rainier Court
Lincoln, NE 68510-5050
Telephone: 402/327-0205
E-Mail: Kurtgeis1@aol.com

EDUCATION

A.B. with Honors, Davidson College
M.S., The University of Georgia
Ph.D., The Pennsylvania State University

ADMINISTRATIVE RESPONSIBILITIES

- 2006 to the present **Director, Buros Center for Testing, University of Nebraska-Lincoln.** Direct the Buros Institute for Mental Measurements. Supervise director of the Buros Institute for Assessment Consultation and Outreach. Provide consultation on assessment issues to clients. Editorial and executive leadership to the **Mental Measurements Yearbooks, Tests in Print**, and the journal, **Applied Measurement in Education**, which I edit. Serve as Interim Director of the Buros Institute for Assessment Consultation and Outreach effective 9/15/2007. Tenured, chaired professor.
- 2001 to 2006 **Vice President for Academic Affairs, University of St. Thomas, Houston**
Responsible for programs encompassing over 125 full-time faculty members and approximately 5,000 students. Lead deans of Schools of Arts & Science, Business, Education, Theology, and Graduate Program in Liberal Arts as well as libraries and advisement. Responsible for personnel, student, and budget issues. Lead the college in the absence of the President. Tenured full professor.
- 1997 to 2001 **Academic Vice President, Le Moyne College.**
Responsible for college of over 120 full-time faculty members and approximately 3,000 students. Lead deans, academic departments, library, admissions, financial aid, registrar, academic support center, and continuing education office. Specific personal responsibility for running graduate programs in Business (MBA) and Education (M.Ed.). Responsible for personnel, student, enrollment and budget issues. Lead the college in the absence of the President. Tenured full professor.

- 1992 to 1997 **Dean of Arts and Sciences, State University of New York, College at Oswego.** Lead 19 academic departments, Biological Field Station, Environmental Research Center, Art Gallery, approximately 20 librarians, Office of Learning Support Services, consisting of over 230 faculty members and 4900 students. Responsible for personnel, student and budget issues. Tenured full professor.
- 1985 to 1991 **Chairperson of the Department of Psychology, Fordham University.** Administered department of 18 full-time faculty members with at least 5 additional FTEs and approximately 180 full- and part-time graduate students, 5 doctoral program, and 70 undergraduate majors. Coordinated extensive faculty evaluation and hiring efforts.
- 1979 to 1985 **Director of Doctoral Program in Psychometrics, Department of Psychology, Fordham University.** Administered doctoral program for approximately 25 graduate students. Wrote descriptions of the program, developed a formal curriculum, coordinated curriculum and course offerings, developed Ph.D. Comprehensive Examinations, advised program students. Coordinated hiring of program faculty.

AWARDS, RECOGNITIONS, AND HONORARY OFFICES

American Board of Assessment Psychology, Diplomate (1994)
Recipient of the Jacob Cohen Award for Distinguished Teaching and Mentoring, American Psychological Association, 2008
Recipient of the President's Award for Scholarly and Creative Activity, SUNY-Oswego, 1995
Recipient of the 1997 Leo D. Doherty Award by the Northeastern Educational Research Association for leadership in educational research
Recipient of the 2002 Thomas J. Donlon Award by the Northeastern Educational Research Association for distinguished mentoring
Biographee in Who's Who in the East (23rd ed., 24th ed., 25th ed., 26th ed., 27th ed., 28th ed.), Who's Who in America (48th ed., 49th ed., 50th ed., 54th ed., 55th ed., 56th ed., 57th ed.) ,Who's Who in American Education (4th ed., 5th ed.), Who's Who in the World (20th ed.), Who's Who in Medicine and Healthcare (3rd ed.), Who's Who in Emerging Leaders in America (4th ed.), Who's Who in Science and Technology
Psi Chi (National Psychology Honor Society)
Sigma Xi (National Scientific Research Society)
Northeastern Educational Research Association, President for term 1987-1988 (President-elect, 1986-87); (Past President, 1988-89)
Northeastern Educational Research Association, Program Committee, 1978 - present (Co-Chair, 1985)
Northeastern Educational Research Association, Member, Board of Directors for term 1984-87
Phi Kappa Phi (National Academic Honor Society)
President, Fordham University Chapter, Phi Kappa Phi, 1985-86
President-Elect and Acting President, Fordham University Chapter, Phi Kappa Phi 1984-85
Treasurer (1983-86) and Secretary (1983-84), Fordham University Chapter, Phi Kappa Phi
Selected as an Outstanding Young Man of America, 1982

FACULTY EXPERIENCE

Faculty Work and Employment

- 2006-present **W. C. Meierhenry Distinguished University Professor, Department of Educational Psychology, University of Nebraska-Lincoln.** Teach one advanced doctoral seminar per semester and run Buros Center for Testing and the Buros Institute for Mental Measurements. Beginning in October 2007, directing the Buros Institute for Assessment Consultation and Outreach on an interim basis. Serve on departmental and university committees.
- 1989 to 1992 **Professor of Psychology, Fordham University, (Tenured).** Chairperson, Graduate School of Arts and Sciences' Long Term Planning Committee (1989-91); Member, University Research Council (1985-1991); Member, Graduate Studies Council (1984-1991); Member, University Tenure Review Committee (1990-1993), Chair (1991-92). Served on various Departmental and University committees. Served as survey and grading consultant to the College Dean. Graduate courses taught included Statistics, Psychological Testing, Test Construction, Psychometrics, Survey and Interview Methodology, Differential Psychology, Personnel Selection, Program Evaluation, the Teaching of Psychology. Undergraduate courses taught: Introductory Psychology, Statistics, Research Design, Psychological Testing, and Seminar on Personnel Decisions for Police. Supervised 16 doctoral dissertations (with one currently in progress). Served on dissertation committees. Supervised Masters' research projects (have directed twelve studies). Advised graduate and undergraduate students. Coordinated faculty evaluation via student rating.
- 1981 to 1989 **Associate Professor of Psychology, Fordham University, (Tenured 6/83).** Essentially the same duties and responsibilities as above. On sabbatical, Spring Semester, 1986, at the Research Division, Educational Testing Service, Princeton, NJ.
- 1977 to 1981 **Assistant Professor of Psychology, Fordham University.** Essentially the same job activities as Professor above.
- 1975 to 1976 **Instructor, Departments of Educational Psychology and Psychology, The Pennsylvania State University.** Taught graduate courses in Educational and Psychological Testing.

Externally Funded Research Activity

- 2006-2013 As Director of Buros Center for Testing at the University of Nebraska, I have brought in approximately \$350,000/year in contract research. One example is listed below.
- 2007-2008 **Project Director.** Department of Education, State of Florida (\$200,000). Provide consultation to the State regarding its statewide testing program, its equating. Discuss implications of testing program with legislators and senators as well as Department of Education commissioners and staff members.
- 1993-94 **Institutional Planning Team Member,** American Council on Education/National Endowment for the Humanities, *Spreading the Word*, a program to institute a Modern Languages across the curriculum project at the State University of New York at Oswego.
- 1988-92 **Faculty Participant and Project Evaluator,** Grant (\$250,000) from the Fund for the Improvement of Post-Secondary Education (FIPSE) to develop a Master of Arts in Liberal Studies at Fordham University.

- 1990 to 1991 **Project Director**, Grant (\$9,840) from the American Psychological Association Science Directorate with matching grant (\$6,560) from the Fordham University Sesquicentennial Celebration to host a conference on the *Psychological Testing of Hispanics*, February 9, 1990, New York City.
- 1979 to 1980 **Project Director**, Grant (\$50,000) from Harcourt Brace Jovanovich, Inc. to study the effects of test use in the school with special emphasis on their use with minority children. Anne Anastasi, Principal Investigator.
- 1978 to 1979 **Research Associate**, Grant (\$75,000) from Harcourt Brace Jovanovich, Inc. Same study as above. Anne Anastasi, Principal Investigator and Project Director.

SERVICE ON NATIONAL COMMITTEES AND BOARDS

- American Educational Research Association, American Psychological Association & National Council on Measurement in Education representative to the International Standards Commission's Committee on International Testing Standards, 2007-2010
- American Psychological Association, Board of Directors, 2011-2013
- American Psychological Association, Council of Representatives member for Division 5 (2006-2010)
- American Psychological Association, Coalition of Academic, Scientific, and Applied Psychology, President-elect (2008), President (2009), Past President (2010)
- American Psychological Association, Committee on Psychological Tests and Assessments, 1998 to the 2000
- American Psychological Association, Committee on Psychological Tests and Assessments Task Force on Test Interpretation for Diverse Groups, 1993 to the 1997
- American Psychological Association, Division 5 (Evaluation, Measurement & Statistics), Member, Membership Committee (1988-1992), Chairperson (1991-92); Member, ad hoc Committee for the Disabled; Member, Public Policy Committee (1996-99), Chairperson (1997-98), Executive Committee Member (2006-2008)
- American Psychological Association, Division 15 (Educational Psychology), Member, Early Contributions Committee (1990-1993)
- American Psychological Association, Office of Program Accreditation and Consultation, Site Visitor, 1987 to 2010
- The College Board, Research Advisory Committee, 2010-2012 (Chair 2012)
- The College Board, Middle States Regional Council, 1998-2000
- The College Board, Member, Editorial Board, The College Board Review, 2000-2006
- The College Board, Scholastic Assessment Test Committee, 2000-2003
- Council for the Accreditation of Educator Professionals, Commission on Standards and Performance Reporting, Commissioner, 2012-2013
- Council for the Accreditation of Educator Professionals, Commission on Institutional Briefs, Commissioner, 2013-present
- Council for the Accreditation of Educator Professionals, Research Committee, Member, 2013-present
- Council for the Accreditation of Educator Professionals, Commission on Standards and Performance Reporting, Commissioner, 2012-2013
- Council of Graduate Schools, Committee on Masters' Education at Predominantly Masters Institutions, 2002- 2006
- Council of Independent Colleges, Committee to Provide a Workshop for New Chief Academic Officers, 2002-2004, Chair (2003-2004)
- Educational Testing Service, Member, Panel convened to review test security procedures and processes (1994)
- Graduate Record Examination, Technical Advisory Committee, Member, 1995-2002; Chair 2000-2003

Graduate Record Examination, Board, *ex officio*, 2000-2003
Graduate Record Examination, Board, 2003-2007
Graduate Record Examination, Chair-elect, 2004-2005, Chair, 2005-2006, Past Chair, 2006-2007
Graduate Record Examination, Research Committee, 2000-2007

International Association of Applied Psychology, Division 2 (Psychological Assessment and Evaluation),
President-elect (2014-2018), President (2018-2022), Past President (2022-2026).

International Test Commission, Council Member (2010-2012), Treasurer (2012-2014)

Joint Committee on Testing Practices, American Psychological Association delegate and Co-Chair (1992-96)
Joint Committee on Testing Practices, Member and Co-Chair, Understanding Testing Working Group, 1990-94
(American Psychological Association delegate to the committee)
Joint Committee on Testing Practices, Member, Testing Individuals with Disabilities Working Group, 1996-2001
Joint Committee on Testing Practices, Member and Co-Chair, Test Taker Rights Working Group, (1993-2001)
National Council on Measurement in Education, Professional Training and Development Committee (1990-92),
Chairperson (1991-92)
National Council on Measurement in Education, Member, Ad Hoc Committee to Develop a Code of Ethical
Standards Committee (1992-94)
National Council on Measurement in Education, Program Committee Co-Chair (1994)

EDITORIAL WORK

2011-2012	Special Issue Editor, International Journal of Testing (Volume 12, Issue 2)
2006 to the present	Editor, Applied Measurement in Education
2001 to the present	Consulting Editor, Practical Assessment, Research, and Evaluation
2000 to the 2006	Consulting Editor, College Board Review
2000 to the present	Consulting Editor, International Journal of Testing
1992 to 2000	Member, Editorial Board, Psychological Assessment
1992 to the present	Member, Editorial Board, Educational Research Quarterly
1997 to the present	Member, Editorial Board, ITEMS
1991 to 1997	Member, Board of Cooperating Editors, Educational and Psychological Measurement
1992 to 1995	Member, Advisory Board, Educational Measurement: Issues and Practice
1988 to 1991	Co-Editor, The NERA Researcher (the quarterly newsletter of the Northeastern Educational Research Association)
1978 to 1983	Consulting Editor, Improving College and University Teaching
1979 to 1984	Consulting Editor, Journal of Educational Research
1988	Consultant, Psychology of Work Behavior (4th ed.), by F. J. Landy. Homewood, IL: Dorsey Press.
1988	Consultant, Psychology (2nd ed.), by L. T. Benjamin, J. R. Hopkins, and J. R. Nation, New York: Macmillan.
1985	Consultant, Psychology: The Science of People (2nd ed.), by F. J. Landy. Englewood Cliffs, NJ: Prentice-Hall, Inc.

- 1983 Consultant and Critical Reader, **Psychology: The Science of People**, by F. J. Landy. Englewood Cliffs, NJ: Prentice-Hall, 1984.
- 1982 Editorial Consultant, **Applied Psychology in Occupational Organizations**, by L. R. Aiken. Reading, MA.: Addison-Wesley.
- 1982 Editorial Consultant, **The Handbook of Questionnaire Construction**, by J. R. Jacoby. New York: Academic Press, 1984.
- 1982 Editorial Consultant, **Principles and Techniques of Questionnaire Design**, by F. J. Kviz and W. L. Kreitman. New York: Academic Press, 1984.
- 1980 Reviewer, **Applied Psychometrics**, by R. L. Thorndike. Boston: Houghton Mifflin Company, 1982.
- 1980 Critical Reader, **Psychology**, by C. Wortman and E. Loftus, New York: Random House, 1981.
- 1980 Editorial Consultant, **Psychology at Work: An Introduction to Industrial Psychology**, by J. P. Houston and L. M. Berry. Boston: Addison-Wesley.
- 1978 Critical Reader, **Psychology Today: An Introduction**, by J. Braun and D. E. Linder. New York: Random House, 1979.

MEMBERSHIPS IN PROFESSIONAL ASSOCIATIONS

American Association for Higher Education
American Educational Research Association
American Psychological Association (Divisions 2 [Teaching of Psychology], 5 [Measurement, Evaluation, Statistics and Assessment], 14 [Industrial/Organizational Psychology] and 15 [Educational Psychology]); Fellow in Divisions 5, 15, and 52
American Psychological Society, Charter Fellow
College Board
Council of Graduate Schools
Eastern Psychological Association
National Council on Measurement in Education
Northeastern Educational Research Association
Northern Rocky Mountain Educational Research Association
Society of Psychologists in Management

SELECTED CONSULTING

- 2008 **Measured Progress.** Serve on a panel to consider the testing of students with disabilities.
- 2008 **The College Board.** Considered validation report of the new SAT.
- 2007 **W.W. Norton** (Publishing house). Served on a panel to make recommendations on improving undergraduate assessment.
- 1995-2003 **Educational Testing Service.** Served on and chaired (2000-2003) the Technical Advisory Committee for the GRE. Served on other paid committees related to the SAT and the GRE.
- 2001, 2002, 2005-6 **Disability Rights Advocates.** Testified before a panel formed to recommend the flagging policies on the SAT and other College Board examinations to the College Board

and against the Association of American Medical Colleges in a case relating to the provision of accommodations to individuals with disabilities.

2002-2003 **U.S. Department of Justice.** Consider the validity of the Law School Admission Test for the Department of Justice and wrote reports regarding the same.

1999 **Stephoe & Johnson, LLP.** Provided a deposition and testimony for the United States District Court of Eastern Pennsylvania regarding the use of flagging for a professional school licensure test.

1986 to the 2001 **Fox and Fox, Counselors at Law,** Newark, NJ. Serve as an expert witness and consultant in cases related to the title of Engineer, Department of Transportation, and Fire Captain. Consulted on other cases related to police and fire matters.

1989 to 1995 **Cornell University, New York State School of Industrial and Labor Relations.** (New York City). Delivered lecture entitled "E.E.O. Selection" several times a year as part of their "Human Resources Programs: Professional Workshop Series in New York City."

1995 to 1998 **American Board of Physical Therapists,** Alexandria, VA. Provide guidance with regard to the passing score for some certification examinations.

Prior to 1995

Prior to 1992, I was engaged in considerable consultation with a variety of states, municipalities, and unions. This consultation generally concerned test development and often involved my leading major test construction projects in civil service, industrial, and educational testing. I gained considerable understanding of the functioning of organizations in so doing.

For approximately 10 years from the early 1980s through the 1990s, I was the primary consultant to **New York City Department of Personnel** building and defending in court Police Officer and Fire Fighter Examinations. This involved examinations for the entry-level positions as well as all levels of promotion including for police service of Sergeant, Lieutenant, and Captain for the New York City Police Department, the Transit Police and the Housing Police. It also included working with all ranks in the Fire Department, through Chief of the Department, as well as positions in Sanitation (entry-level and promotional), Social Services, Parks and Recreation, and Health Services. I built the examination used for the hiring of Test and Measurement Specialists in the New York City.

I served as an expert witness (**working with the New York City Department of Law**) in a number of cases defending New York City personnel examinations and in one case, against the examination for Parks and Recreation Worker. I served as an expert witness and consultant to the **Public Service Alliance of Canada**, the union of federal employees in Canada, in cases against three tests, a Canadian Intelligence Test, a Canadian Customs Officer Supervisor Test, and a Office Manager Test.

Together with Dr. Richard R. Reilly of Assessment Alternatives, I performed job analyses for the New Jersey Civil Service Commission of a number of police positions.

While in graduate school at the Pennsylvania State University, for almost two years I directed a court-ordered study that ultimately brought women onto the police force in the City of Philadelphia. Prior to and during this study, I went through various aspects of police training, worked closely with uniformed police representatives of the department, developed rating scales for the evaluation of police officers. As stated above, this was a full-time position for approximately 1.5 years with Bartell Associates, of State College, Pennsylvania.

RESEARCH ACTIVITY AND PUBLICATIONS

Articles for Journals

- Dahlman, K. A. & Geisinger, K. F. (2015, in press). The Prevalence of Measurement in Undergraduate Psychology Curricula across the United States. *Scholarship of Teaching and Learning in Psychology*.
- Lee, H. & Geisinger, K. F. (In press.) The Matching Criterion Purification for DIF Analyses in a Large-scale Assessment. *Educational and Psychological Measurement*.
- Brabeck, M. M., Dwyer, C. A., Geisinger, K. F., Marx, R. W., Noell, G. W., Pianta, R. C., Subotnick, R. F., & Worrell, F. C. (2015, in press). Assessing the assessments of teacher preparation. *Theory into Practice*, DOI: : 10.1080/00405841.2015.1036667
- Lee, H. & Geisinger, K.F. (2014). The Effect of Propensity Scores on DIF Analysis: Inference on the Potential Cause of DIF. *International Journal of Testing*, 14, 313-338
- Geisinger, K. F. (2014). Established best practices. *NCME Newsletter*, 22(4), p. 6.
- Geisinger, K. F. (2012). Worldwide test reviewing at the beginning of the twenty-first century. *International Journal of Testing*, 12, 103-107.
- Carlson, J. F. & Geisinger, K. F. (2012). Test reviewing at the Buros Center for Testing. *International Journal of Testing*, 12, 122-135.
- Geisinger, K. F., Kriegsman, K., Leigh, I. W., Manghi, E., Schultz, I.Z., Seekins, T., & Taliaferro, T. (Authorship by this committee and listed in alphabetical order.) (2012). Guidelines for assessment of and intervention with persons with disabilities. *American Psychologist*, 67, 43-62.
- Patel, N. P., Bussler, J. F., Geisinger, K. R., Geisinger, & Hill, I. D. (2011). Are Pathologists Accurately Diagnosing Eosinophilic Esophagitis in Children? A 9Year Single Academic Institutional Experience With Interobserver Observations. *International Journal of Surgical Pathology*, 19, 290-296.
- Geisinger, K. F. (2010). Consequences and validity: An uneasy relationship. *NCME Newsletter*, 18 (1; March, 2010), 8-9.
- Geisinger, K. F. & McCormick, C. M. (2010). Adopting Cut Scores: Post-Standard-Setting Panel Considerations for Decision Makers. *Educational Measurement: Issues and Practice*, 29 (1; March, 2010), 38-44.
- Geisinger, K. F. (2010). Report on the 7th Conference of the International Test Commission. *International Psychology Bulletin*, 2010, 14 (4), 48-49.
- A College Admissions Question: What would we do if the SAT and ACT did not exist. *NCME newsletter*, 17 (2; June 2009), 4-6.
- Invited essay: Some reflections on faculty evaluation. *Newsletter of the Society for Teaching of Psychology*. (Spring, 2009). 4-5.
- Screening: Testing the Limits. *Human Resource Executive*, 2008, 17-19. Also available at: <http://www.hreonline.com/HRE/story.jsp?storyId=146568445>.
- The Revised GRE General Test launch in Fall 2007. (K. F. Geisinger & D. Payne). *CGS Communicator*, March, 2006, XXXIX (2), pp. 4.

Coming soon to your campus: The Revised GRE General Test and the TOEFL IBT. . (With D. G. Payne & R. Setzler). CGS Communicator, November, 2005, XXXVIII (9), pp. 4-5.

Improving the Graduate Admissions Process: How Deans Can Influence Program Decision Making. CGS Communicator, July, 2004, XXXVII, pp. 1,2,5,7.

Psychological testing at the end of the millennium: A brief historical review. Professional Psychology, 2000, 31, 117-118.

Scholarship in psychology: A paradigm for the 21st Century, American Psychologist, 1998, 54, 1292-1297. (With D. F. Halpern, D. Smothergill, M. Allen, S. Baker, C. Baum, D. Best, J. Ferrari, E. Gilden, M. Hester, P. Keith-Spiegel, N. C. Kierniesky, T. V. McGovern, W. J. McKeachie, W. F. Prokasy, C. T. Szuchma, R. Vasta, and K. A. Weaver.)

Using subject matter experts to assess content representation: An MDS analysis. (With S. G. Sireci.) Applied Psychological Measurement, 1995, 19, 241-255.

Responding to graduate students' professional deficiencies: A nationwide survey. (With M. E. Procidano, N. Busch-Rossnagel, and M. Reznikoff.) Journal of Clinical Psychology, 1995, 51, 426-433.

Development and preliminary validation of the Ego Identity Process Questionnaire. Journal of Adolescence. 1995, 18, 179-192. (With E. Balistreri and N. Busch-Rossnagel.)

Psychometric issues in testing students with disabilities. Applied Measurement in Education, 1994, 7, 121-140.

Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychological Assessment, 1994, 6, 304-312.

Psychometric and assessment issues raised by the Americans with Disabilities Act (ADA). (With M. L. Tenopyr, W. H. Angoff, J. N. Butcher, and R. R. Reilly.) The Score, 1993, 15 (4), pp. 1-2, 7-15.

Analyzing test content using cluster analysis and multidimensional scaling. (With S. G. Sireci.) Applied Psychological Measurement, 1992, 16, 17-31.

The metamorphosis in test validation. Educational Psychologist, 1992, 27, 197-222.

Using standard setting data to establish operational cutoff scores. Educational Measurement: Issues and Practice, 1991 10 (2), 17-22.

The metamorphosis in test validation. The NERA Researcher, 1991, 29 (1), 2-12.

Response latency to computer administered inventory items as an indicator of emotional arousal. (With D. E. Temple.) Journal of Personality Assessment, 1990, 54, 289-297.

The relationship between recruiting source, applicant quality, and hire performance. (With J. Powell-Kirnan and J. A. Farley.) Personnel Psychology, 1989, 42, 293-308.

The Golden Rule in psychological testing: Please, please don't do it unto me. Theoretical and Philosophical Psychology, 1988, 8, 15-22.

Presidential address: Legal issues in test construction, validation and use. The NERA Researcher, 1988, 26 (4), 2-7.

The psychosocial impact and development implications of the threat of nuclear war on adolescents. (With D. Rudoy and M. Reznikoff.) Medicine and War, 1987, 3, 77-91.

Psychosocial development and stressful life events among religious professionals. (With S. D. Sammon and M. Reznikoff.) *Journal of Personality and Social Psychology*, 1985, 48, 676-687.

Cross-validation of the factor structure of the McGill Pain Questionnaire. (With M. Byrne, A. Troy, L. A. Bradley, P. J. Marchisello, L. H. Van der Heide, and E. J. Prieto.) *Pain*, 1982, 13, 193-201.

The prediction of graduate school success in psychology. (With J. Powell-Kirnan.) *Educational and Psychological Measurement*, 1981, 41, 815-820.

The use of tests with schoolchildren: Final project report. (With A. Anastasi.) *Journal Supplement Abstract Service*, 1981, 11, 58. Also abstracted in *Resources in Education*, March, 1981.

A construct validation of faculty orientations toward grading: Comparative data from three institutions. (With A. N. Wilson and J. J. Naumann.) *Educational and Psychological Measurement*, 1980, 40, 413-417.

The language of low back pain: Factor structure of the McGill Pain Questionnaire. (With E. J. Prieto, L. Hopson, L. A. Bradley, M. Byrne, O. Midax, and P. J. Marchisello.) *Pain*, 1980, 8, 11-19.

Serum chloride: A College of American Pathologists' survey. (With K. R. Geisinger, P. Wakely, and J. G. Batsakis.) *American Journal of Clinical Pathology*, 1980, 74, 546-551.

Individual differences among college faculty members in grading. (With W. Rabinowitz.) *Journal of Instructional Psychology*, 1980, 7, 20-27.

Who are giving all those A's? An examination of high grading college faculty members. *The Journal of Teacher Education*, 1980, 31 (March-April), 11-15.

A note on grading policies and grade inflation. *Improving College and University Teaching*, 1979, 27, 113-115.

Book Chapters and Entries in Edited Works

Geisinger, K.F. & Usher-Tate, B. J. (In press). Face validity. In A. Wenzel (Ed.), *The Sage Encyclopedia of Abnormal and Clinical Psychology*. Thousand Oaks, CA: Sage.

Geisinger, K.F. (In press). Incremental validity. In A. Wenzel (Ed.), *The Sage Encyclopedia of Abnormal and Clinical Psychology*. Thousand Oaks, CA: Sage.

Geisinger, K.F. (In press). Inter-rater reliability. In A. Wenzel (Ed.), *The Sage Encyclopedia of Abnormal and Clinical Psychology*. Thousand Oaks, CA: Sage.

Geisinger, K.F. (In press). Reliability. In A. Wenzel (Ed.), *The Sage Encyclopedia of Abnormal and Clinical Psychology*. Thousand Oaks, CA: Sage.

Geisinger, K.F. (In press). Validity. In A. Wenzel (Ed.), *The Sage Encyclopedia of Abnormal and Clinical Psychology*. Thousand Oaks, CA: Sage.

Geisinger, K.F. (In press). Test construction. In J. Norcross & VandenBos, G. (Eds), *APA Handbook of Clinical Psychology*. Washington, DC: APA Books.

Geisinger, K.F. (In press). Technology and Test Administration: The Search for Validity. In F. Drasgow (Ed.), *Technology and Testing: Improving Educational and Psychological Measurement*. Washington, DC: NCME.

- Geisinger, K. F. (In press). A Brief Review of Spanish-Language Adaptations of Some English-Language Intelligence Tests. In K. F. Geisinger (Ed.), *Psychological Testing of Hispanics: Clinical, Cultural, and Intellectual Issues*. Washington, DC: APA.
- Geisinger, K. F. & McCormick, C. (In press). Testing individuals with disabilities: An international perspective. In Leong, F. T. L., Bartram, D., Cheung, F.M., Geisinger, K. F., Hattie, J.A., & Iliescu, D. (Eds.) *International Test Commission Handbook of Testing*, Oxford, Eng.: Oxford University Press.
- Geisinger, K. F. & Usher-Tate, B. J. (In press). The History of Educational Testing and Psychometrics. In C. S. Wells & Falkner-Bond (Eds.) *Educational Measurement: From Foundations to Future*. NY: Guilford.
- Geisinger, K. F. (In press). Test evaluation. In Lane, S., Raymond, M., & Haladyna, T. Test evaluation. In Lane, S., Raymond, M., & Haladyna, T. (Eds.), *Handbook of test development (2nd ed.)*. Washington, DC: NCME.
- Geisinger, K. F. (2013). Reliability. In K. F. Geisinger (Ed.), *Handbook of testing and assessment in psychology* (vol. 1; pp. 21-42). Washington, DC: American Psychological Association.
- Geisinger, K. F. (2013). Review of the "Adult Basic Learning Examination, Second Edition." In Wood, C. & Hays, D. G., (Eds.), *A counselor's guide to career assessment instruments* (pp. 121-124). Broken Arrow, OK: National Career Development Association.
- Folers, D., Cotner, H., Kotamraju, P., & Geisinger, K. F. (2012). Using data for decision-making, accountability and evaluation. In D. Folkers, K. Green, R. Hinkley, & D.Mills (Eds.), *The career pathways effect: Linking education and economic prosperity* (pp. 301-331). Waco, TX: CORD Communications.
- Geisinger, K. F. & McCormick, C. Testing and assessment in cross-cultural psychology. (2012). In Naglieri, J., & Graham, J. (Eds.), *Handbook of Psychology, Volume 10: Assessment Psychology* (pp. 161-224). NY, NY: Wiley.
- Geisinger, K. F. (2012.) Norm- and criterion-referenced testing. In H. Cooper (Ed.), *Handbook of research methods in psychology* vol. 1, pp. 371-393). Washington, DC: American Psychological Association.
- Geisinger, K. F., Shaw, L. H., & McCormick, C. (2011). The validation of tests in higher education. In C. Secolsky & D. B. Denison (Eds.), *The Handbook of Measurement, Assessment and Evaluation in Higher Education* (pp. 194-207). NY, NY: Routledge.
- Geisinger, K. F. (2011). The future of high-stakes testing in education. In J. A. Bovaird, K. F. Geisinger, & C. W, Buckendahl (Eds.). *High Stakes Testing: Science and Practice in K-12 Settings* (pp. 231-248). Washington, DC: APA Books.
- Bovaird, J. A., Geisinger, K. F. & Buckendahl, C. W. (2011). Introduction. In J. A. Bovaird, K. F. Geisinger, & C. W, Buckendahl (Eds.). *High Stakes Testing: Science and Practice in K-12 Settings* (pp. 3-10). Washington, DC: APA Books.
- Geisinger, K. F. (2010). Opening comments. Included in presentations from the Buros Center Conference for Monitoring Assessment Quality in the Age of Accountability, The eighteenth mental measurements yearbook (pp. 807-809). Lincoln, NE: Buros Institute of Mental Measurements.
- Geisinger, K. F. (2010). Closing thoughts: A look to the future. Included in presentations from the Buros Center Conference for Monitoring Assessment Quality in the Age of Accountability, The eighteenth mental measurements yearbook (pp. 863-868). Lincoln, NE: Buros Institute of Mental Measurements.
- Review of Adult Basic Learning Examination (Second Edition). (2009). In E. A. Whitfield, R. Feller, & C. Wood (Eds.), *A Counselor's Guide to Career Assessment Instruments* (5th ed.) (pp. 90-93). Broken Arrow, OK: National Career Development Association.

- Bogardus Social Distance Scale. (2010). In I. Weiner & W. E. Craighead (Eds.) Corsini encyclopedia of psychology (4th ed.). (Vol. 1; p. 246.) New York: Wiley.
- Psychometrics: Norms, Reliability, Validity, and Item Analysis. (2010). In I. Weiner & W. E. Craighead (Eds.) Corsini encyclopedia of psychology (4th ed.). (Vol. III; pp. 1345-1349). New York: Wiley.
- Questionnaires. (2010). In I. Weiner & W. E. Craighead (Eds.) Corsini encyclopedia of psychology (4th ed.). (Vol. III; pp. 1408-1410). New York: Wiley.
- Test standardization. (2010). In I. Weiner & W. E. Craighead (Eds.) Corsini encyclopedia of psychology (4th ed.). (Vol. IV; pp. 1769-1770). New York: Wiley.
- Testing methods. (2010). In I. Weiner & W. E. Craighead (Eds.) Corsini encyclopedia of psychology (4th ed.). (Vol. IV; pp. 1773-1774). New York: Wiley.
- Foreword. (2009.) In G. Robertson, L. Eyde, & S. Krug. Responsible test use: Case studies for assessing human behavior (2nd ed.) (pp. xi-xii.) Washington, DC: American Psychological Association.
- The future of high stakes testing. (In press.) In J. Bovaird, K. F. Geisinger & C. B. Buckendahl (Eds.), High stakes testing. Washington, DC: American Psychological Association.
- Geisinger, K. F., & Carlson, J. F. (2009). Standards and standardization. In J. N. Butcher (Ed.), Oxford handbook of personality assessment (pp. 99-111). New York, NY: Oxford University Press. [updated; an earlier version was published in 2002 in J. N. Butcher (Ed.), Clinical personality assessment: Practical approaches (2nd ed., pp. 243-256). New York, NY: Oxford University Press.
- Psychological diagnostic testing. (2008.) In R. Phelps (Ed.) The anti-testing fallacies (pp. 67-88). Washington, DC: American Psychological Association.
- General Aptitude Test Battery. (2008). In F. Leong (Ed.), Encyclopedia of Counseling Volume 4: pp. 1541-1542). Thousand Oaks, CA: Sage.
- Preface. (2007). In C. Calahan Laitusis & L.L. Cook (Eds.), Large Scale Assessment and Accommodations: What Works? (pp. ix-xii). Alexandria, VA: Council on Exceptional Children.
- The testing industry, ethnic minorities, and those with disabilities. (2005). In R. Phelps (Ed.), Defending standardized testing (pp. 187-203). Mahwah, NJ: Erlbaum.
- Conversion of the Wechsler Adult Intelligence Scale into Spanish: An early test adaptation effort of considerable consequence. (With C. Y. Maldonado.) (2005). In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), Adapting educational and psychological tests for cross-cultural assessment (pp.213-234). Mahwah, NJ: Erlbaum.
- Review of the *College Student Experiences Questionnaire* (2005). In Spies, R. A. & Plake, B. S. (Eds.) The sixteenth mental measurements yearbook. The Buros Institute on Mental Measurements, University of Nebraska-Lincoln, Lincoln, NE. Pp. 244-247.
- Review of the *Power and Performance Measures* (2005). In Spies, R. A. & Plake, B. S. (Eds.) The sixteenth mental measurements yearbook. The Buros Institute on Mental Measurements, University of Nebraska-Lincoln, Lincoln, NE. Pp. 804-807.
- Bogardus Social Distance Scale. (2004.) In W. E. Craighead & C. B. Nemeroff (Eds.), Corsini concise encyclopedia of psychology and behavioral science (3rd ed.). New York: Wiley. Pp. 131-132.

- Psychometrics: Norms, Reliability, Validity, and Item Analysis. (2004.) In W. E. Craighead & C. B. Nemeroff (Eds.), Corsini concise encyclopedia of psychology and behavioral science (3rd ed.). New York: Wiley. Pp. 758-760.
- Questionnaires. (2004.) In W. E. Craighead & C. B. Nemeroff (Eds.), Corsini concise encyclopedia of psychology and behavioral science (3rd ed.). New York: Wiley. Pp. 787-788.
- Testing methods. (2004). In W. E. Craighead & C. B. Nemeroff (Eds.), Corsini concise encyclopedia of psychology and behavioral science (3rd ed.). New York: Wiley. Pp. 983-984.
- Test standardization. (2004.) In W. E. Craighead & C. B. Nemeroff (Eds.), Corsini concise encyclopedia of psychology and behavioral science (3rd ed.). New York: Wiley. Pp. 984-986.
- Testing students with Limited English Proficiency. (2004). In J. E. Wall & G. R. Walz (Eds.), Measuring up: Assessment issues for teachers, counselors, and administrators (147-159). Greensboro, NC: ERIC Clearinghouse for Counseling and Student Services and the National Board for Certified Counselors.
- Testing and Assessment in Cross-Cultural Psychology. (2003). In J. R. Graham & J. A. Naglieri (Eds.), Handbook of Psychology (Volume 10: Assessment Psychology) (95-117). I. B. Weiner (Editor-in-Chief). New York: John Wiley.
- The psychometrics of testing individuals with disabilities. (With G. Boodoo & J. P. Noble). (2002). In R. Ekstrom & D. K. Smith (Eds.), Assessing Individuals with Disabilities in Educational, Employment, and Counseling Settings. (pp. 33-42). Washington, DC: American Psychological Association.
- Standards and standardization. (With J. F. Carlson.) (2002). In J. N. Butcher (Ed.), Clinical Personality Assessment: Practical approaches. (2nd ed.; pp. 243-256). New York: Oxford University Press.
- Testing the members of an increasingly diverse society. (2002). In J. F. Carlson & B. B. Waterman (Eds.), Social and Personality Assessment of School-Aged Children: Developing Interventions for Educational and Clinical Use. Boston: Allyn & Bacon. Pp. 346-364.
- Development of a statement of Test Taker Rights and Responsibilities. (2001). In G. R. Walz & J. C. Bleuer (Eds.), Assessment: Issues and Challenges for the Millennium. Greensboro, NC: CAPS Publications/ERIC Clearinghouse for Counseling & Student Services. Pp. 143-162.
- Testing students with disabilities. (With J. F. Carlson). (2001). In G. R. Walz & J. C. Bleuer (Eds.), Assessment: Issues and Challenges for the Millennium. Greensboro, NC: CAPS Publications/ERIC Clearinghouse for Counseling & Student Services. Pp. 375-380.
- Review of the *Wonderlic Personnel Inventory/Scholastic Level Examination*. (2001). In J. Impara, (Ed.) Buros Mental Measurements Yearbook: Volume XIV. Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln. Pp. 1360-1363.
- Review of the *Skillscan for Management Development*. (2001.) In J. Impara, (Ed.) Buros Mental Measurements Yearbook: Volume XIV. Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln. Pp. 1142-1146..
- Bogardus Social Distance Scale. (2001.) In W. E. Craighead & C. B. Nemeroff (Eds.), The Corsini encyclopedia of psychology and behavioral science (3rd ed.). New York: Wiley. Vol. 1, p. 225.
- Psychometrics: Norms, Reliability, Validity, and Item Analysis. (2001.) In W. E. Craighead & C. B. Nemeroff (Eds.), The Corsini encyclopedia of psychology and behavioral science (3rd ed.). New York: Wiley. Vol 3, pp. 1313-1316.

- Questionnaires. (2001.) In W. E. Craighead & C. B. Nemeroff (Eds.), The Corsini encyclopedia of psychology and behavioral science (3rd ed.). New York: Wiley. Vol 4, pp. 1362-1364.
- Test standardization. (2001.) In W. E. Craighead & C. B. Nemeroff (Eds.), The Corsini encyclopedia of psychology and behavioral science (3rd ed.). New York: Wiley. Vol. 4. pp. 1683-1684.
- Testing accommodations for the new millennium: Computer-administered testing in a changing society. In Niyogi, S. (Ed.). New Directions in Assessment for Higher Education: Fairness, Access, Multiculturalism & Equity (FAME). The Graduate Record Examination FAME Report Series, No. 2, (1998), pp. 12-20.
- Equity issues in employment testing. (With S. Sireci). In J. Sandoval, C. Frisby, K. F. Geisinger, J. Scheuneman, & J. M. Ramos-Grenier (Eds.) Test interpretation and diversity: Achieving equity in psychological assessment. Washington, DC: American Psychological Association, 1998, pp. 105-140.
- Psychometric issues involved in test interpretation. In J. Sandoval, C. Frisby, K. F. Geisinger, J. Scheuneman, & J. M. Ramos-Grenier (Eds.) Test interpretation and diversity: Achieving equity in psychological assessment. Washington, DC: American Psychological Association, 1998, pp. 17-30.
- Training psychologists to assess members of a diverse society. (With J. F. Carlson). In J. Sandoval, C. Frisby, K. F. Geisinger, J. Scheuneman, & J. M. Ramos-Grenier (Eds.) Test interpretation and diversity: Achieving equity in psychological assessment. Washington, DC: American Psychological Association, 1998, pp. 375-386.
- Review of the *Tests of Adult Basic Education Work-Related Foundation Skills (TABE-WRFS)*. In J. C. Impara & B.S. Plake (Eds.), The Thirteenth Mental Measurements Yearbook. Lincoln, NE: Buros Institute of Mental Measurements, 1998, pp. 1086-1088.
- Review of the *Watson-Glaser Critical Thinking Appraisal, Form S*. In J. C. Impara & B.S. Plake (Eds.), The Thirteenth Mental Measurements Yearbook. Lincoln, NE: Buros Institute of Mental Measurements, 1998, pp. 1121-1124.
- Review of the *National Police Officer Selection Test*. In J. J. Kramer & J. C. Conoley (Eds.), The Twelfth Mental Measurements Yearbook. Lincoln, NE: Buros Institute of Mental Measurements, 1995, pp. 672-675.
- Review of the *NOCTI Teacher Occupational Competency Test*. In J. J. Kramer & J. C. Conoley (Eds.), The Twelfth Mental Measurements Yearbook. Lincoln, NE: Buros Institute of Mental Measurements, 1995, pp. 682-685.
- Standards and standardization. (With J. F. Carlson.) In J. N. Butcher (Ed.), Clinical Personality Assessment: Practical approaches (pp. 211-223). New York: Oxford University Press, 1995.
- Testing students with disabilities. (With J. F. Carlson.) ERIC/Clearinghouse on Counseling and Student Services Digest, 1995, (ERIC #EDO-CG-95-27). In W. D. Schafer (Guest Ed.), Assessment in counseling and therapy: An ERIC/CASS Special Digest Collection, Washington, DC: ERIC Clearinghouse on Counseling and Student Services.
- Psychometric and policy issues in the use of tests with individuals with disabilities. Proceedings of the Joint Conference on Disability Issues, pp. 141-145. April, 1995.
- Testing LEP students for minimum competency and high school graduation. In Focus on Evaluation and Measurement (Volume 2). Washington, DC: United States Department of Education, Office of Bilingual Education and Minority Languages Affairs, 1992, pp. 33-67.
- Review of the Management Competence Index. In J. J. Kramer & J. C. Conoley (Eds.), The Eleventh Mental Measurements Yearbook. Lincoln, NE: Buros Institute of Mental Measurements,

- 1992, pp. 502-503. Also available as Accession Number AN-11180749, Mental Measurements Yearbook Database (Search Label MMYD), BRS Information Technologies.
- Review of the PSB-Nursing School Aptitude Examination (RN). In J. J. Kramer & J. C. Conoley (Eds.), The Eleventh Mental Measurements Yearbook. Lincoln, NE: Buros Institute of Mental Book Measurements, 1992, pp. 719-721. Also available as Accession Number AN-11180749, Mental Measurements Yearbook Database (Search Label MMYD), BRS Information Technologies.
- Assessing language-minority students. (With J. F. Carlson.) ERIC/Clearinghouse on Tests, Measurement and Evaluation Digest, 1992, EDO-TM-92-4.
- Fairness and selected psychological issues in the psychological testing of Hispanics. In K. F. Geisinger (Ed.), Psychological testing of Hispanics. Washington, D.C.: American Psychological Association, 1992, pp. 17-42.
- Preface. In K. F. Geisinger (Ed.), Psychological testing of Hispanics. Washington, D.C.: American Psychological Association, 1992, pp. xv-xvii.
- Bogardus Social Distance Scale. In R. J. Corsini (Ed.), Concise encyclopedia of psychology. New York: Wiley, 1987, pp. 146-147.
- Psychometrics. In R. J. Corsini (Ed.), Concise encyclopedia of psychology. New York: Wiley, 1987. pp. 925-926.
- Questionnaires. In R. J. Corsini (Ed.), Concise encyclopedia of psychology. New York: Wiley, 1987, pp. 952-953.
- Test Standardization. In R. J. Corsini (Ed.), Concise encyclopedia of psychology. New York: Wiley, 1987. p. 1115.
- The General Aptitude Test Battery. (With J. Kirnan.) In R. C. Sweetland & D. J. Keyser (Eds.), Test Critiques: Volume V. Kansas City: Test Corporation of America, 1986. pp. 150-167. Reprinted in Keyser, D. J., & Sweetland, R. D. (Eds.), Test critiques compendium: Reviews of major tests from the Test Critiques series. Kansas City: Test Corporation of America, 1987. pp. 163-180. Reprinted in Bolton, B., (Ed.), Special education and rehabilitation testing: Current practices and test reviews. Austin, TX: Pro-ed, 1988, pp. 217-234.
- The Miller Analogies Test. In R. C. Sweetland & D. J. Keyser (Eds.), Test critiques: Volume III. Kansas City: Test Corporation of America, 1985, pp. 414-424.
- The ACT Assessment. In R. C. Sweetland & D. J. Keyser (Eds.), Test critiques: Volume I. Kansas City: Test Corporation of America, 1985, pp. 11-28.
- Bogardus Social Distance Scale. In R. J. Corsini (Ed.), Wiley encyclopedia of psychology. New York: Wiley, 1984. Volume I, p. 160.
- Psychometrics. In R.J. Corsini (Ed.), Wiley encyclopedia of psychology. New York: Wiley, 1984. Volume 3, pp. 163-165.
- Questionnaires. In R. J. Corsini (Ed.), Wiley encyclopedia of psychology. New York: Wiley, 1984. Volume 3. pp. 199-200.
- Test standardization. In R.J. Corsini (Ed) Wiley encyclopedia of psychology. New York: Wiley, 1984. Volume 3. p. 414.
- Factor analytic studies of the McGill Pain Questionnaire. (With E. J. Prieto). In R. Melzack (Ed.), Pain

measurement and assessment. New York: Raven Press, 1983. pp. 63-70.

Marking systems. In H. E. Mitzel (Ed.), Encyclopedia of educational research (5th Ed.) New York: The Free Press, 1982. Vol. 3: 1139-1149.

Grading attitudes and practices among college faculty members. (With W. Rabinowitz.) In H. Dahle, A. Lysne, & P. Rand (Eds.), A Spotlight on educational problems. Oslo, Norway: Universitets Forlaget, 1979. pp. 145-172. (Distributed in the United States by the Columbia University Press, Irvington, NY).

Developing an operational model for assessing experiential learning. (With W. W. Willingham.) In W. W. Willingham & H. S. Nesbitt (Eds.), Implementing a program for assessing experiential learning. Princeton, NJ: Cooperative Assessment of Experiential Learning (Educational Testing Service), 1976. Chapter 1; pp. 1-15.

Overview of CAEL field research. (With W. W. Willingham.) In W. W. Willingham & Associates, The CAEL validation report, Princeton, NJ: Cooperative Assessment of Experiential Learning (Educational Testing Service), 1976. Chapter III, pp. 1-35.

Data analysis. (With R. R. Reilly & W. W. Willingham.) In W. W. Willingham & Associates, The CAEL validation report, Princeton, NJ: Cooperative Assessment of Experiential Learning (Educational Testing Service), 1976. Appendix 5, pp. 5-1 - 5-42.

Books and Monographs

- Pardes, H., Barsky, A., Daly, M., Geisinger, K. F., Gerber, N., Jette, A., Koop, J. Suzuki, L. A., Twamley, E., Ubel, P., & Wall, J. (2015, in press). *Psychological testing in the service of disability determination*. Washington, DC: National Academies Press (Institute of Medicine report/to be published by the Institute of Medicine).
- Geisinger, K. F. (2015, in press). (Ed.) *Psychological testing of Hispanics: Clinical, cultural, and intellectual assessment*. Washington, DC: American Psychological Association.
- Carlson, J. F., Geisinger, K. F. & Jonson, J. (Eds.) (2014) *The Nineteenth Mental Measurements Yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.
- Worrell, F. C., Brabeck, M. M., Dwyer, C. A., Geisinger, K. F., Marx, R. W., Noell, G. H., & Pianta, R.C. (2014). *Assessing and evaluating teacher preparation programs: An APA Task Force Report*. Washington, DC: American Psychological Association.
- Schlueter, J. E., Carlson, J. F., Geisinger, K. F., and Murphy, L.L. (Eds.) (2013). *Pruebas Publicadas en Español: An index of Spanish tests in print*. Lincoln, NE: Buros Center for Testing.
- Geisinger, K. F. (Ed.). (2013). *Handbook of testing and assessment in psychology* (3volumes). Washington, DC: American Psychological Association.
- Murphy, L. M., Geisinger, K. F., Carlson, J. F., & Spies, R. S. (2011). *Tests in print VIII*. Lincoln, NE: Buros Institute of Mental Measurements.
- Bovaird, J., Geisinger, K. F., & Buckendahl, C. B. (Eds.), (2011.) *High Stakes Testing: Science and Practice in K-12 Settings*. Washington, DC: American Psychological Association.
- Spies, R. A., Carlson, J. F., & Geisinger, K. F. (Eds.) (2010). *The Eighteenth Mental Measurements Yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.
- Geisinger, K. F., Spies, R. A., Carlson, J.F., & Plake, B. S. (2007). *The Seventeenth Mental Measurements Yearbook*. Lincoln, NE: Buros Institute for Mental Measurements.
- Sandoval, J., Frisby, C., Geisinger, K.F., Scheuneman, J., & Ramos-Grenier, J. M. (Eds.). (1998). *Test interpretation and diversity: Achieving equity in psychological assessment*. Washington, DC: American Psychological Association.
- Lloyd, B., Crocker, L., Geisinger, K.F., & Webb, M. (1994). *Report of the panel convened to review test security procedures at the Educational Testing Service in February, 1994*. Princeton, NJ: Educational Testing Service.
- Geisinger, K. F. (1992). *Psychological Testing of Hispanics* (Ed.), Washington, DC: American Psychological Association, 1992.
- Geisinger, K. F. & Anastasi, A. *Instructor's manual to accompany Psychological testing*, A. Anastasi, Sixth Edition. New York: Macmillan, 1988.
- Geisinger, K. F. & Anastasi, A. *Instructor's manual to accompany Psychological testing*, A. Anastasi, Fifth Edition. (With A. Anastasi and S. Urbina.) New York: Macmillan, 1982.
- Anastasi, A. & Geisinger, K. F. Use of tests with schoolchildren *JSAS Catalogue of Selected Documents in Psychology*, 1981 11, ERIC Document No. ED 194-635.

Videotapes

The ABC's of School Testing. (Project Co-director, with J. J. Fremer and J. Wall). Co-author of the Leader's Guide for the videotape. Both are available from the National Council on Measurement in Education, Washington, DC. (1994).

Book Reviews

Review of The Conditions of Admission: Access, Equity, and the Social Contract of Public Universities by John Aubrey Douglas. Educational Horizons, 2008, 86, 182-185.

One is the loneliest number: Two is not as bad as one (in some instances). (With S. L. Davis.) Review of Dyadic Data Analysis by D. A. Kenny, D. A. Kashy, W. L. Cook. Journal of Clinical and Social Psychology, 2008, 27, 311-313.

Review of Making sense of college grades. (By O. Milton, H. R. Pollio, and J. A. Eison.) Journal of Educational Measurement, 1988, 25, 167-170.

Review of Support for teaching at major universities. (Edited by S. C. Ericksen with J. A. Cook.) Improving College and University Teaching, 1980, 28, 41.

Paper Presentations

Geisinger, K. F. (2015). The ITC Guidelines on Quality Control in Scoring, Test Analysis and Reporting of Test Scores. In A. Odendall (Chr.), The International Test Commission's Guidelines for Good Testing Practice. Symposium presented at the annual meeting of the Society for Industrial and Organizational Psychology, Philadelphia, PA. April.

Geisinger, K. F. (2015). Using ITC Guidelines. In D. Bartram (Chr.), Executive Board Special Session: Improving International Testing Practice with the International Test Commission. Symposium presented at the annual meeting of the Society for Industrial and Organizational Psychology, Philadelphia, PA. April.

Geisinger, K. F. (2015). Global transportability of measures. In Y. Yang & T. L. Hayes (Co-chairs), Transportability: Boundaries, Challenges, and Standards. Symposium presented at the annual meeting of the Society for Industrial and Organizational Psychology, Philadelphia, PA. April.

Geisinger, K. F. (2015). Publishing in Applied Measurement in Education. Roundtable presented at the annual meeting of the American Educational Research Association, Chicago, IL. April.

Geisinger, K. F. (2015). General Overview of Standards for Technical Quality. In Worrell, F. (Chr.), Higher Education Assessment: Evaluating and Assessing Teacher Preparation Programs. Symposium presented at the annual meeting of the American Educational Research Association, Chicago, IL, April.

Geisinger, K. F. (2015). Test reviewing at the Buros Center for Testing. In T. Patelis (Chr.), Various Efforts to Evaluate the Quality of Assessment Programs. Symposium presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL, April.

Geisinger, K. F. (2015). The assessment of 21st Century skills: A global perspective. Invited address at the University of Luxembourg, Luxembourg. March 4, 2015.

Geisinger, K. F. (2014). Keynote interview. In Kristen Huff, (Chair). Invited keynote presentation at the annual meeting of the Northeastern Educational Research Association, October, Trumbull, CT.

- Geisinger, K. F. (2014). The Buros Approach to Ensuring Quality. In T. Patelis (Chr.)/Ensuring the quality of assessments. Symposium presented at the annual meeting of the Northeastern Educational Research Association, October, Trumbull, CT.
- Geisinger, K. F. (2014). How do we ensure fairness? In T. Patelis (Chr.), Fairness issues in assessment and accountability. . Symposium presented at the annual meeting of the Northeastern Educational Research Association, October, Trumbull, CT.
- Geisinger, K. F. (2014). Evaluating tests: A continuing effort for psychologists. Invited divisional keynote presentation, International Congress of Applied Psychology, Paris, France, July, 2014.
- Geisinger, K. F. (2014). International technical standards for test quality and test reviewing. In D. Bartram (Chair), Symposium presented at the International Congress of Applied Psychology, Paris, France, July, 2014..
- Geisinger, K. F. (2014). Assessing 21st Century Skills. Invited workshop presented at the biannual meeting of the International Test Commission, San Sebastian, Spain, July, 2014.
- Geisinger, K. F. (2014). Preparing doctoral-level psychometrics specialists. In T. Oakland (Chair), How do we prepare psychometric specialists. Symposium presented at the biannual meeting of the International Test Commission, San Sebastian, Spain, July, 2014.
- Geisinger, K. F. (2014). Applied Measurement in Education. Roundtable with a journal editor presented at the annual meeting of the American Educational Research Association, Philadelphia, PA, April, 2014.
- Lee, H. & Geisinger, K. F. (2014). Differential item functioning analysis models in large-scale assessment. Paper editor presented at the annual meeting of the American Educational Research Association, Philadelphia, PA, April, 2014.
- Lee, H. & Geisinger, K. F. (2014). Purification of the matching criterion in the equated pooled booklet method for DIF. Paper editor presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA, April, 2014
- Geisinger, K. F. (2013). Outcomes assessment. Lecture presented at King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, November.
- Geisinger, K. F. (2013). Classroom assessment. Workshop presented at King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, November.
- Geisinger, K. F. (2013). Best practices for faculty in graduate admissions. Workshop presented at King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, November.
- Geisinger, K. F. (2013). Grading: Assessment technique and learning facilitator. Workshop presented at King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, November.
- Geisinger, K. F. (2013). Setting the minimum passing score for the CFA examination. Workshop presented to the Chartered Financial Analyst Board of Governors, London, England, November.
- Geisinger, K. F. (2013). Tensions between Educational/Political Realities and Reliability and Validity. In F. Worrell (Chair), Effective use of data for program improvement. Symposium presented at the annual meeting of the American Psychological Association, Honolulu, August.
- Geisinger, K. F. (2013). Building unbiased assessments. Workshop presented to the faculties of the Bryan College of Health Sciences, Clarkson College, and Nebraska Methodist College.

- Lee, H.S. & Geisinger, K. F. (2013). Efficiency of Generalized Full Information Bifactor Model. Poster presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April.
- Geisinger, K. F. (2013) Contributions of Anne Anastasi. In S. Sinharay (Chair), A look at our psychometric history: Contributions of Thurstone, Lindquist, Anastasi, Bock, Messick, and Holland. Symposium presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April.
- Geisinger, K. F. (2013). The future of admissions testing in the United States. Invited keynote at the Buros "Big Issues in Testing Conference." Lincoln, NE, March.
- Geisinger, K. F. (2013). A testing course focusing on diversity issues. In K. F. Geisinger (Chair), Making a quantitative program more multicultural. Symposium presented at the National Multicultural Summit and Conference, Houston, January.
- Geisinger, K. F. The future content in admissions testing. Invited presentation to the First International Conference on Assessment & Evaluation: Admissions Criteria in Higher Education, Riyadh, Saudi Arabia, December, 2012.
- Geisinger, K. F. (2012). Criterion-referenced testing. Paper presented at the Ronference Honoring Professor Ronald Hambleton, Amherst, MA, November.
- Geisinger, K. F., Carlson, J. F., & Jonson, J. Evaluating tests: Fundamental concepts and skills for psychologists and researchers. Continuing Education Workshop presented at the annual meeting of the American Psychological Association, August, 2012.
- Geisinger, K. F. & Bartram, D. International Perspectives on Test Reviewing. Paper presented at the Quadrennial meeting of the International Congress of Psychology, Cape town, SA, July, 2012.
- Geisinger, K. F. Evaluating tests: Fundamental concepts and skills for psychologists and researchers. Workshop presented at the biannual meeting of the International Test Commission, Amsterdam, the Netherlands, July, 2012
- Geisinger, K. F. Languages and linguistic diversity. In P. Elosua (Chair), Linguistic diversity and testing. Symposium presented at the biannual meeting of the International Test Commission, Amsterdam, the Netherlands, July, 2012.
- Geisinger, K. F. The testing of multiple languages in a single country. In . In D. Sandilands (Chair), Assessment of linguistic minority students in Canada and the United States. Symposium presented at the biannual meeting of the International Test Commission, Amsterdam, the Netherlands, July, 2012.
- Geisinger, K. F. Some thoughts on international test adaptations. In J-L. Padilla (Chair), Challenges of test adaptation in special contexts: The role of the ITC Guidelines. Symposium presented at the biannual meeting of the International Test Commission, Amsterdam, the Netherlands, July, 2012.
- Carlson, J. F. & Geisinger, K. F. (2012) Test reviewing at the Buros Center for Testing. In K. F. Geisinger (Chair), International Perspectives on Test Reviewing. Symposium presented at the biannual meeting of the International Test Commission, Amsterdam, the Netherlands, July, 2012.
- Geisinger, K. F. (2012). A 50,000 foot view on observed score equating. In M. Wiberg (Chair), New developments in observed score equating. Symposium presented at the biannual meeting of the International Test Commission, Amsterdam, the Netherlands, July, 2012.
- McCormick, C. M., Shaw, L. H., Evers, A., & Geisinger, K. F. (2012). A multilevel approach to the EFPA/ITC questionnaire on test attitudes. In A. Evers (Chair), Attitude of psychologists on tests and testing: The

results of an international survey. Symposium presented at the biannual meeting of the International Test Commission, Amsterdam, the Netherlands, July, 2012.

Geisinger, K. F. Cultural bias in testing. Presentation at the First Session of the Summer Faculty and Staff Development Series, BryantLGH College of Health Sciences, May, 2012.

Geisinger, K. F. Anne Anastasi's Views on Ability and Achievement. Invited paper presented at the Hertz Memorial Presentation in Memory of Anne Anastasi at the annual meeting of the Society for Personality Assessment, Chicago, IL, March, 2012.

Geisinger, K. F. & Shaw, L. H. Evaluation of Accuplacer®, PSAT/NMSQT, and SAT program features. Presentation to the Research Division of the College Board, New York City (also broadcast to Newtown, PA). March, 2012.

Geisinger, K. F. & Patelis, T. Maintenance schedules for quality. Presentation to the Research Advisory Committee of the College Board, Phoenix, AZ, March, 2012.

Geisinger, K. F. The future of admissions testing. Invited presentation at the ETS Conference on the Future of Learning, Education and Assessment, Educational Testing Service, Princeton, NJ, March, 2012.

Geisinger, K. F. The scholarly and fair evaluation of psychological tests and assessments: English language and adapted tests. Invited workshop at the First Caribbean Regional Conference on Psychology, Nassau, Bahamas, November, 2011.

Geisinger, K. F. Testing and psychometrics at NERA. In R. Michel (Chair). Designing statewide testing programs. Symposium presented at the annual meeting of the Northeastern Educational Research Association, Hartford, CT, October 2011.

Geisinger, K. F. If we could change K-12 testing today. In T. Patelis (Chair). Past presidents discuss educational research. Symposium presented at the annual meeting of the Northeastern Educational Research Association, Hartford, CT, October 2011.

Geisinger, K. F. Change and stability: Revisiting new recurrent concerns. In K. F. Geisinger, (Chair). Issues in large scale testing. Symposium presented at the annual meeting of the Northeastern Educational Research Association, Hartford, CT, October 2011.

Geisinger, K. F. Diversity and psychometrics: A necessary but almost null hypothesis. In Diversity in Psychometrics, P. Scott-Johnson (Chr.), Symposium presented at the annual meeting of the American Psychological Association, Washington, DC, August, 2011.

Geisinger, K. F. Test reviewing at the Buros Center for Testing. In D. Bartram (Chr.), Internationalization of test reviewing. Symposium presented at the biannual meeting of the European Congress of Psychology, Istanbul, Turkey, July, 2011.

Geisinger, K. F. Validation: Its role in Test Reviews at the Buros Center for Testing. In S. Sireci (Chr.), Validating educational and psychological tests; Theory, applications, and future directions. Symposium presented at the biannual meeting of the European Congress of Psychology, Istanbul, Turkey, July, 2011.

Byrne, B. M., Geisinger, K. F. & Oakland, T. The work of the International Test Commission. Symposium presented at the Fifth Brazilian Congress of Assessment Psychology, Bento Goncalves, Brazil, June 2011.

Geisinger, K. F. Scientific Publication in Psychological Assessment: Challenges toward the internationalization of the knowledge. In E. Remor (Chr.), Scientific Publication in Psychological Assessment: Challenges toward the internationalization of the knowledge. Symposium presented at the Fifth Brazilian Congress of Assessment Psychology, Bento Goncalves, Brazil, June 2011.

- Geisinger, K. F. The Scholarly Evaluation of Tests and Assessments. Invited keynote address presented at the Fifth Brazilian Congress of Assessment Psychology, Bento Goncalves, Brazil, June 2011.
- Chin, T. Y., Geisinger, K. F. & Yang, Y. (2011). Classification Accuracy of Diagnostic Methods: A Simulation Study. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA, April, 2011.
- Geisinger, K. F. (2011). The history of the Buros Center for Testing. In T. Patelis (Chr.). Perspectives on the history of testing in the United States. Symposium presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA, April, 2011.
- Geisinger, K. F. (2011). Some thoughts on the breadth of educational and psychological testing. Invited lecture at the University of Kansas, Lawrence, KS. March, 2011.
- Geisinger, K. F. The Buros Center for Testing and its Admissions Testing Initiatives. Invited paper presented to the students of the SRAM program, UNL, Lincoln, NE, November 2010.
- Geisinger, K. F. Classical test theory. In Graduate Students Issues Committee Special Invited Session on Advanced Measurement and Statistics. Symposium presented at the annual meeting of the Northeastern Educational Research Association, Hartford, CT, October, 2010.
- Geisinger, K. F. Alternate assessment: Should assessment drive instruction? Paper presented at the annual meeting of the Northeastern Educational Research Association, Hartford, CT, October, 2010.
- Geisinger, K. F. Dissertations: Hurdles, pathways or gateways. In T. Patelis (Chr.) On Finishing and Further: Dissertation Research Now and Then. Symposium presented at the annual meeting of the Northeastern Educational Research Association, Hartford, CT, October, 2010.
- Geisinger, K. F. Reviewing tests: A comprehensive approach. Presentation to the College Board Research and Development staff, Newtown, PA, October, 2010.
- Geisinger, K. F. Reviewing manuscripts for Applied Measurement in Education. Presentation to the College Board Research and Development staff, Newtown, PA, October, 2010.
- Geisinger, K. F. Consequences and validity. In E. Burke (Chr.) Reconsidering Messick: Validity and best practices in testing. Symposium presented at the biennial meeting of the International Test Commission, Hong Kong, July 2010.
- Geisinger, K. F. An American Psychometrician's Perspective. In M. Ph. Born (Chr.), Informing about ISO 10667- An International Standard for Assessment Service Delivery in Work and Organizational Settings. Symposium presented at the biennial meeting of the International Test Commission, Hong Kong, July 2010.
- Geisinger, K. F. Evaluating Test Quality as Users and Writing Manuals as Authors: Two Sides of a Coin. Workshop presented at the biennial meeting of the International Test Commission, Hong Kong, July 2010.
- Geisinger, K. F. College Admissions Testing for Student Selection: Challenges for Deans and Vice President. Paper presented to the Hochschulrektorskonferenz (Conference of University Presidents). Bonn, Germany, December, 2009.
- Geisinger, K. F. College admissions testing at German universities: What might such testing look like and what advantages and disadvantages might it bring? Paper presented to the RWH (University of Aachen) Psychology Department, December 2009.
- Geisinger, K. F. Concepts of validity. In Patelis, T., Conceptions of validity. Symposium presented at the annual meeting of the Northeastern Educational Research Association, Hartford, CT, October, 2009.

Geisinger, K. F. How to get published. In K. Huff (Chair), Symposium for New Faculty Members. Symposium presented at the annual meeting of the Northeastern Educational Research Association, Hartford, CT, October, 2009

Geisinger, K. F. Testing Issues and Concerns. Invited presentation to the Career and Technical Education State Collaborative Working Group of the Council of Chief State School Officers. Baltimore, MD, October 2009.

Geisinger, K. F. Paper-and-pencil vs. Computer-based Test Delivery. Invited presentation to the Career and Technical Education State Collaborative Working Group of the Council of Chief State School Officers. Baltimore, MD, October 2009.

Geisinger, K. F. A College Admissions Question: What would we do if the SAT and ACT did not exist? Paper presented at the annual meeting of the Northern Rocky Mountain Educational Research Association, Jackson Hole, October, 2009. (Also presented to the QQPM Seminar at the University of Nebraska-Lincoln, November, 2009).

McCormick, C. M. & Geisinger, K. F. When do testing accommodations give an unfair advantage? A Comparison to a double-amputee sprinter's quest to compete in the Olympics. Paper presented at the annual meeting of the Northern Rocky Mountain Educational Research Association, Jackson Hole, WY, October, 2009.

Foley, B.P., Geisinger, K.F., Roschewski, P., & Foy, E. (2009, October). *Conducting an alignment study in the context of a performance assessment with a single writing prompt*. Paper presented at the Annual meeting of the Northern Rocky Mountain Educational Research Association, Jackson Hole, WY.

Geisinger, K. F. Non-Traditional Admissions Measures in Higher Education: Some Comments. In P. Kyllonen (Chr.), New constructs and new measures in higher education admissions. Symposium presented at the annual meeting of the American Psychological Association, Toronto, CA, August, 2009.

Geisinger, K. F. Research on the SAT-Writing Test. Discussant Comments in W. Camara (Chair), The SAT Writing Test: An Update on Research. Toronto, CA, August, 2009.

The Buros Institute of Mental Measurements Test Review Process (With J. F. Carlson.) In D. Bartram (Chr.) Symposium on national approaches to test quality assurance. Symposium presented at the 11th Biannual European Congress of Psychology, Oslo, NO, July, 2009.

Status update on the revision of the US Joint Standards on Testing. In E. Burke (Chr.), International guidelines and standards related to tests and testing. Symposium presented at the 11th Biannual European Congress of Psychology, Oslo, NO, July, 2009.

An educational testing perspective on the ITC testing quality control guidelines in scoring, analysis and reports. In A. Allalouf & M. Born (Co-Chrs.) The development of ITC guidelines on quality control in scoring, analysis and reports. Symposium presented at the 11th Biannual European Congress of Psychology, Oslo, NO, July, 2009.

Eosinophilic Esophagitis: Interobserver Variability in a Disease Entity in Which Counting Counts. (With J. F. Busler, N. Patel, I.D. Hill, & K.R. Geisinger). Poster presented at the annual meeting of the United States and Canadian Academy of Pathology, Boston, March 2009.

American Psychological Association Science Agenda Goals. (With M. L. Cooper). Workshop presented to the Coalition of Academic, Scientific, and Applied Research Psychologists, American Psychological Association Building, Washington, DC, Feb. 19, 2009.

The Buros Center for Testing at the University of Nebraska. Invited address at the University of Aachen, Aachen, Germany, December, 2008.

- Adjusting standards to enhance validity: Post standard-setting panel considerations to enhance validity. (With C. McCormick.) In K. Huff (Chr.), Validating Standards on Educational Tests, Symposium presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT, October 2008.
- A focus and follow-up on fairness. In T. Patelis (Chr.), The Fordham Five's Fundamentals of Fairness. Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT, October 2008.
- Testing issues and concerns: An introductory presentation to the State Directors of Career and Technical Education. Invited Keynote address to the biannual meeting of the State Directors of Career and Technical Education, Mystic, CT, October, 2008.
- Three significant roles in the teaching of measurement: Mentor, administrator, and campus consultant. Jacob Cohen Award Speech invited at the annual meeting of the American Psychological Association, Boston, MA, August, 2008.
- Issues in cross-cultural testing: The future was yesterday. In B. Byrne (Chr.) Interplay of cross-cultural comparisons and related methodological practices. Symposium presented at the annual meeting of the American Psychological Association, Boston, MA, August, 2008.
- The rights and responsibilities of test takers and test makers. Invited keynote address to the biannual meeting of the International Testing Commission, Liverpool, Eng., July, 2008.
- Anne Anastasi's views on ability and achievement: Implications for the training of measurement professionals. In T. Patelis (Chr.), The legacy of Anne Anastasi on educational research and assessment: Commemorating the 100th anniversary of her birth. Symposium presented at the annual meeting of the American Educational Research Association, New York NY, March, 2008.
- Current validation practice for academic achievement tests. (With C. McCormick & A. Römhild.) Paper presented at the annual meeting of the National Council on Measurement in Education, New York NY, March, 2008.
- Timeliness in meeting the testing standards. In T. Patelis (Chr.), Maintaining quality in large-scale assessment (a.k.a. Maintenance Schedules: They're not just for your car.) Symposium presented at the annual meeting of the Association of Test Publishers, Dallas, TX, March, 2008.
- An international standard (ISO) for assessment in work and organizational settings. (With D. Bartram & W. Camara.) Symposium presented at the annual meeting of the Association of Test Publishers, Dallas, TX, March, 2008.
- The historical and present role of the Buros Institute of Mental Measurements. In K. F. Geisinger (Chr.) Test evaluation in the 21st Century. Symposium presented at the annual meeting of the Association of Test Publishers, Dallas, TX, March, 2008.
- From the Bronx through New Brunswick to Lincoln, Nebraska: Critical Questions in the Review of Tests. (With J. F. Carlson). Paper presented at the annual meeting of the Northeastern Educational Research Association, Hartford, CT, October, 2007.
- Implications of the Spellings Commission for Outcomes Assessment in Higher Education. Paper presented at the annual meeting of the Northern Rocky Mountain Educational Research Association Meeting, Jackson Hole, WY, October 2007.
- Assessment after the Spellings Commission. Paper presented as an after-dinner presentation to the Academic Leadership Dinner held at the annual meeting of the American Psychological Association, San Francisco, CA, August, 2007.

- Improving test use. In L. Stricker (Chair), *Improving Test Use*. Discussant comments presented at a symposium presented at the annual meeting of the American Psychological Association, San Francisco, CA, August, 2007.
- Changes in the Verbal Test of the GRE. In K. F. Geisinger (Chair), *The Revised Graduate Record Examination General Test—Requisite Knowledge*. Paper presentation in a symposium presented at the annual meeting of the American Psychological Association, San Francisco, CA, August, 2007.
- The future of high stakes testing. Keynote address at the Barbara Plake Festschrift Celebration, Lincoln, NE, May, 2007.
- Investigating students with disabilities on the SAT. In D. L. Morgan (Chair), *Investigating students with disabilities on the SAT*. Symposium presented at the annual meeting of the American Educational Research Association, Chicago, IL, April, 2007.
- Non-cognitive predictors and academic success. In A. E. Schmidt (Chair), *The use of non-cognitive measures for guidance and selection*. Discussant comments presented at an invited symposium presented at the annual meeting of the American Psychological Association, New Orleans, LA, August, 2006.
- Changes in Large-Scale Admissions Measures in American Higher Education: Implications for Test Adaptation. (With D. G. Payne). Invited paper presented at the fourth biannual conference of the International Test Commission, Brussels, BE, July, 2006.
- The New GRE Test. In D. G. Payne (Chair), *The New GRE General Test and GRE 2005 Volume Report*. Symposium presented at the annual meeting of the Council of Graduate Schools, Palm Springs, CA, December, 2005.
- The New GRE Test. (With D. Piacentino.), *The New GRE General Test*. Paper presented at the annual meeting of the Association of Texas Graduate Schools, Lubbock, TX, October, 2005.
- The New GRE Test. In D. G. Payne (Chair), *The New GRE General Test and GRE 2004 Volume Report*. Symposium presented at the annual meeting of the Council of Graduate Schools, Washington, DC, December, 2004.
- Development of a Statement of Test Taker Rights and Responsibilities. In N. Abeles (Chair), *Ethical Issues in Assessment*. Invited symposium presented at the annual meeting of the American Psychological Association, Honolulu, HI, August, 2004.
- Revisions to the GRE General Test. In D. Johnson (Chair), *Use of the GRE and the Analytic Writing Measure in Master's and Ph.D. Programs: Views from the Field*. Symposium at the annual meeting of the Council of Graduate Schools, San Francisco, December, 2003.
- An Update on the Graduate Record Examination. Presentation at the annual meeting of the Association of Texas Graduate Schools, San Angelo, TX, September, 2003.
- An Administrative Perspective on Part-Time Faculty Members: The Issue of Best Utilizing Adjuncts. In J. F. Carlson (Chair), *Don't quit your day job: Perspectives on Part-Time Teaching*. Symposium presented at the annual meeting of the American Psychological Association, Toronto, ON, August 2003.
- Psychometric Issues in Testing Individuals with Disabilities: Instructional Validity. Invited Keynote Symposium entitled "High Stakes Testing: Challenges, Victories and Best Practices" at the annual meeting of the International Dyslexia Association, Atlanta, GA, November, 2002.
- Some thoughts on Dr. Thomas F. Donlon, My Friend and Mentor. Acceptance remarks upon receipt of the Thomas F. Donlon Award, presented at the annual meeting of the Northeastern Educational Research Association, Kerhonksen, NY, October, 2002.

Anne Anastasi's continuum of experiential specificity for tests of developed ability and the current SAT controversy. Paper presented in the Tribute to Anne Anastasi Symposium at the American Psychological Association, Susana Urbina (Chair), Chicago, IL, August, 2002.

Some language issues in educational and psychological testing. Paper presented at the annual meeting of the American Psychological Association, San Francisco, August, 2001.

Some Thoughts on the Matter of Flagging: Reactions to a Trial. Paper presentation to the annual meeting of the National Council on Measurement in Education, Seattle, WA, April, 2001.

Some issues in the college use of Advanced Placement tests. (With D. DePerro.) Invited presentation to the annual meeting of the Middle States Regional Council of the College Board, Baltimore, MD, February, 2000.

Testing individuals who do not fit the mold. Invited presentation to the Psychometrics program, University of Massachusetts, Amherst, MA, October, 1999.

Considerations in adapting intelligence tests: A focus on the Wechsler Tests. Invited presentation at the Joint European Conference of the International Association for Cross-Cultural Psychology and the International Test Commission., Graz, Austria, June, 1999.

A review of some Spanish-language adaptations of some English-language intelligence tests. Keynote address presented at the International Conference on Test Adaptation: Adapting Tests for Use in Multiple Languages and Cultures, Washington, DC, May, 1999.

Psychometric issues in achieving equity in psychological assessment. In J. Sandoval (Chair), Test interpretation and diversity: Achieving equity in psychological assessment. Symposium presented at the annual meeting of the American Psychological Association, San Francisco, CA, August, 1998.

Some Summative Thoughts on Sternberg's Paper and the Validity of the Graduate Record Examination in Graduate Admissions. In A. R. Fitzpatrick (Chair), Evaluating the predictive validity of the Graduate Record Examination. Symposium presented at the annual meeting of the American Psychological Association, San Francisco, CA, August, 1998.

An interprofessional project on rights and responsibilities of test takers. In H.E. Roberts-Fox (Chair), Test-taker rights and responsibilities: Issues and perspectives. Symposium presented at the annual meeting of the American Psychological Association, San Francisco, CA, August, 1998.

Faculty use of the GRE in graduate admissions: What is the validity? Paper presented at the annual meeting of the Northeastern Association of Graduate Schools, Baltimore, MD, April, 1998. (Also presented to the Technical Advisory Committee for the Graduate Record Examination, Educational Testing Service, Princeton, NJ, June, 1998)

The Library of the future: One academic administrator's reflections. Keynote address presented at the annual meeting of the New York State Library Assistant's Association, Syracuse, NY, June, 1998.

A brief history of test taker rights and responsibilities: A call for codification. In J. Noble, (Chair), The rights and responsibilities of test takers. Symposium presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA, April, 1998.

Psychometric issues involved in test interpretation for members of diverse groups. In H. Roberts-Fox (Chair), Test interpretation and diversity: Achieving equity in assessment, Invited symposium presented at the Assessment '98: Assessment for change—Changes in assessment conference, St. Petersburg, FL, January, 1998.

A multi-profession project to enumerate the rights and responsibilities of test takers. In K. F. Geisinger & W. Schafer (Co-chairs), Test taker rights and responsibilities, Invited symposium presented at the Assessment '98: Assessment for change—Changes in assessment conference, St. Petersburg, FL, January, 1998.

Pathways to organizational diversity in the next millennium: Observations of a personnel testing specialist and a college administrator. Invited address to the International Training Conference on Public Personnel Administration: Human Resource Management—Stepping out of the Box. Doris T. McGuffey (Session Chair). Minneapolis, MN, September, 1997.

Suggestions for improving test adaptation practice: Discussant comments. In H. Swaminathan (Chair), Large scale test adaptation projects: Designs, results, and suggestions for improving practice. Symposium presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL, March, 1997,

Testing accommodations for a new millennium: Computer administered testing for a changing society. Invited paper presented at the Invitational Conference on Testing and Higher Education, co-sponsored by Educational Testing Service and Xavier University, New Orleans, March, 1997.

Development of a statement of test takers' rights and responsibilities: Implications for Counselors. In R. Ekstrom, (Chair), The Work of the Joint Committee on Testing Practices. Invited Symposium at the annual meeting of the American Counseling Association, Orlando, FL, March, 1997.

Selected measurement contributions of Harold E. Mitzel. In M. E. Horan, (Chair), A Tribute to Harold E. Mitzel: A founder of NERA and a leader in educational research. Symposium presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY, October, 1996,

Development of a statement of test takers' rights and responsibilities. Paper presented as Introductory Remarks to the Open Conference on Test Taker Rights and at the national headquarters of the American Speech Language Hearing Association, Rockville, MD, October, 1996.

The civil service testing of Hispanics. Invited presentation to the Personnel Testing Council of Metropolitan Washington, Washington, DC, September, 1996.

Advances in test adaptation. Discussant comments. In W. Camara (Chair), Adapting and translating educational and psychological tests: Issues, technical advances, and guidelines. Symposium presented at the annual convention of the American Psychological Association, Toronto, CA, August 1996.

Testing people who do not fit the mold. Presidential Scholarly and Creative Activity Award Address presented at the annual Quest conference, Oswego, NY, April, 1996.

The rights of test takers. Paper presented at the California Test Bureau/McGraw-Hill, Monterey, CA, February, 1996. Also presented at Educational Testing Service, Princeton, NJ, June, 1996.

The rights of test takers: A brief history. In K. F. Geisinger, (Chair), The rights of test takers. Symposium presented at the annual meeting of the American Speech Hearing Language Association, Orlando, FL, December, 1995.

The Joint Committee on Testing Standards. In S. Goldsmith, (Chair), The ABC's of School Testing: A Videotape. Symposium presented at the annual meeting of the American Speech Hearing Language Association, Orlando, FL, December, 1995.

The development of a statement of test taker rights. In W. D. Schafer (Chair), Test taker rights. Symposium presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April, 1995.

Reactions from a member of the development committee. In C. B. Schmeiser, (Chair), Making the ideal real: Dissemination and use of the NCME Code of Ethics. Symposium presented at the annual meeting of the

National Council on Measurement in Education, San Francisco, April, 1995.

Psychometric and policy issues in the use of tests with individuals with disabilities. Paper presented at the Joint Conference on Disability Issues sponsored by the American Bar Association, the Association of American Law School, the Law School Admission Council, and the National Conference of Bar Examiners, St. Louis, MO, April, 1995.

A consideration of graduate education. In V. Hall, (Chair), Graduate education in psychology. Symposium presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY, October, 1994.

Needed changes in the Revised Standards for Educational and Psychological Testing. In W. J. Camara (Chair), Revision of the Standards for Educational and Psychological Testing. Symposium presented at annual conference of the American Psychological Association, Los Angeles, CA, August, 1994.

A summary of four reviews of the NCME Code of Professional Responsibility in Educational Assessment. In C. B. Schmeiser (Chair), Membership Forum on the Proposed NCME Code of Ethics. Symposium presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA, April, 1994.

Who exactly are the testing police? In W. C. Camara (Chair), Enforcing Professional Standards in Measurement (or Do We Need the Testing Police?). Symposium presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA, April, 1994.

The Work of the Joint Committee on Testing Standards: The ABC's of School Testing. In D. K. Smith (Chair), The ABC's of School Testing: A video for parents. Invited symposium presented at the annual meeting of the National Association of School Psychologists, Seattle, WA, March, 1994.

The NCME Code of Professional Responsibility in Educational Assessment: Its development and orientation. In K. F. Geisinger (Chair), Reactions to the NCME Code of Ethical Assessment Practices in Education. Symposium presented at the annual conference of the Northeastern Educational Research Association, Ellenville, NY, October, 1993.

The study of psychological testing of Hispanics: A beginning with a focus on industrial applications. Address presented to the SUNY Oswego chapter of Sigma Xi, Oswego, NY, September, 1993.

Two SUNY-Oswego teacher education partnerships. In J. E. Milley (Chair), "Renewing Partnerships," Symposium presented at the Teach America II: Implementing Teacher Education Reform Conference, Washington, DC, June, 1993.

Functions and uses of the Code of Ethical Assessment Practices in Education. In C.B. Schmeiser (Chair), Ethics in Educational Assessment. Symposium presented at the Council of Chief State School Offices 1993 National Conference on Large Scale Assessment, Assessment: Key to Systematic Change, Albuquerque, NM, June, 1993.

Standards in standardization: United we stand. Keynote address presented at the annual spring seminar of the Counseling and Psychological Department, Oswego, NY, April, 1993.

Ethics in the professions: The case of educational assessment. Invited keynote address at the Phi Kappa Phi Initiation Ceremony, Fordham University, New York, NY, April, 1993.

Audiences, functions and uses of the Code of Ethical Assessment Practices in Education. In C. B. Schmeiser (Chair), NCME Code of Ethics: Reactions to a draft. Symposium presented at the annual meeting of the National Council on Education, Atlanta, GA, April, 1993.

Using subject matter experts to assess content representation: An MDS analysis. (With S. G. Sireci.) Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA, April, 1993.

Perspectives on research on teacher education. Paper presented at the annual convention at the Northeastern Educational Research Association, Ellenville, NY, October, 1992.

Initial validation of placement examinations at a community college. (With D. G. Seguin & K. S. Sweeney.) Paper presented at the annual convention of the Northeastern Educational Research Association, Ellenville, NY, October, 1991.

The psychological testing of Hispanics in industry. Paper presented at the monthly meeting of the Connecticut Applied Psychological Association, New Haven, CT, September, 1991.

Testing LEP students for minimum competency and graduation. Commissioned paper for the National Research Symposium on Limited English Proficient (LEP) Students' Issues: Focus on Evaluation and Measurement, Washington, DC, September, 1991.

Disclosing interpreted test scores to test takers: What are the problems? In J. C. Hansen (Chair), Understanding test results: What should users and examinees know? Symposium presented at the American Psychological Association, San Francisco, CA, August, 1991.

The graduate admissions process in psychology. Psi Chi Invited Lecture presented at the annual meeting of the Eastern Psychological Association, New York, NY, April, 1991.

The metamorphosis in test validation. Invited address presented at the annual convention of the Northeastern Educational Research Association, Ellenville, NY, November, 1990.

Selecting and evaluating a site for the annual convention. Paper presented at the annual convention of the American Educational Research Association, San Francisco, CA, March, 1989.

Using standard setting data to establish operational cutoff scores. In B. H. Loyd (Chair), Practical issues in conducting a standard setting study. Symposium presented at the annual convention of the National Council on Measurement in Education, San Francisco, CA, March, 1989.

Legal issues in test construction, validation and use. Presidential address presented at the annual convention of the Northeastern Educational Research Association, Ellenville, NY, November, 1988.

Post-hoc strategies for insuring and improving content validity. Paper presented as part of a symposium entitled, Issues related to content validation for selection of municipal employees, at the annual convention of the International Personnel Management Association Assessment Council, Philadelphia, PA, July, 1987.

Whither educational research? Roundtable presented at the annual meeting of the Northeastern Educational Research Association, Kerhonksen, NY, October, 1986.

Grading non-cognitive student behavior: A construct validation. (With V. W. Hevern, S. J.) Paper presented at the annual meeting of the American Psychological Association, Washington, DC, August, 1986.

The impact of the 1985 Joint Testing Standards on civil service testing. Invited address as part of the Visiting Scholar Lecture Series, Department of Personnel, New York, NY, March, 1986.

The relationship and stability of two item-bias detection indices. (With G. Locke.) Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, April, 1985.

- The microcomputer as a research tool: Statistical packages. Pre-session presented at the annual convocation of the Northeastern Educational Research Association, Ellenville, NY, October, 1984.
- A questionnaire approach to college curriculum evaluation. Paper presented at the annual convocation of the Northeastern Educational Research Association, Ellenville, NY, October, 1984.
- Ethnic group differences in the personal biserial index. (With F. J. Breyer). Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April, 1984.
- Public personnel selection testing and the law. Invited address as part of the Visiting Scholar Lecture Series, New York City Department of Personnel, New York, NY, April, 1984.
- An initial classification of non-cognitive student behavior grading items. (With V. W. Hevern, S.J.) Paper presented at the annual meeting of the American Psychological Association, Anaheim, CA, August, 1983. ERIC Document No. [PS 014-211](#).
- The relationships of attitudes toward multiple-choice tests and convergent production, divergent production, and risk-taking. (With D. T. Horber.) Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada, April, 1983. ERIC Document No. ED 229-435.
- Sex: A moderator variable between sex role and statistics performance. Paper presented at the annual meeting of the Eastern Psychological Association, Philadelphia, PA, April, 1983.
- Can scientific thinking be measured? (With P. Biesmeyer & H. Koritz.) Paper presented at the annual convention of the National Science Teachers Association and the Society of College Science Teachers, Dallas, TX, April, 1983.
- Construct validation of faculty orientations toward grading: An experimental investigation of differential grade assignment. (With G. Locke.) Paper presented at the annual convocation of the Northeastern Educational Research Association, Ellenville, NY, October, 1982.
- A validation of the Veterinary Aptitude Test. Paper presented at the annual convocation of the Northeastern Educational Research Association, Ellenville, NY, October, 1981.
- Development of a scale to measure attitudes toward multiple-choice testing. (With D. Horber.) Paper presented at the annual convocation of the Northeastern Educational Research Association, Ellenville, NY, October, 1981.
- Cross-validation of the factor structure of the McGill Pain Questionnaire. (With L. A. Bradley, M. Byrne, . Troy, L. Hopson Van der Heide & E.J. Prieto.) Paper presented at the annual meeting of the Eastern Psychological Association, New York, NY, April, 1981.
- Grade inflation and the potential for discrimination in graduate admissions. (With D. Grudzina and M. A. Glynn.) Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles, CA, April, 1981.
- The differential prediction of graduate school success for experimental and clinical psychology students. (With J. Powell-Kirnan.) Paper presented at the annual meeting of the Eastern Educational Research Association, Philadelphia, PA, March, 1981.
- A factor analysis of teachers' attitudes about standardized testing. Paper presented at the annual convocation of the Northeastern Educational Research Association, Ellenville, NY, October, 1981.
- The incremental validity of an MMPI underachievement scale in predicting academic performance. (With

T.J. Dignelli.) Paper presented at the annual meeting of the Eastern Psychological Association, Hartford, CT, April, 1980.

The language of low back pain: Factor structure. (With L. Hopson, E.J. Prieto, L. A. Bradley, & M. Byrne.) Paper presented at the annual meeting of the Eastern Psychological Association, Hartford, CT, April, 1980.

An MMPI underachievement scale as a predictor of academic achievement among high school students. (With V. Hevern, S.J.) Paper presented at the annual meeting of the American Educational Research Association, Boston, MA, April, 1980.

Faculty techniques for preventing cheating: Some baseline data. (With J. J. Maiorca & J. J. Naumann.) Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA, April, 1980.

Intra-university variations in grading: A rationale for differing standards. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA, April, 1980.

Grading and the psychology of motivation. Invited address to the National Conference on Higher Education, Washington, DC, March, 1979.

Faculty orientations toward grading at three academic institutions. (With A. N. Wilson & J. J. Naumann.) Paper presented at the annual convocation of the Northeastern Educational Research Association, Ellenville, NY, October, 1979.

Academic policy and faculty-related changes influencing grading standards. In K. F. Geisinger (Chair), University grade inflation: Documentation, causes, and consequences. Symposium presented at the annual meeting of the American Psychological Association, New York, NY, September, 1979.

Individual differences among college faculty in awarding grades. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA, April, 1979.

Grading policies and grade inflation. Paper presented at the annual convocation of the Northeastern Educational Research Association. Ellenville, NY, October, 1978.

Individual differences in calculator attitudes and performance in a statistics course. (With D. M. Roberts.) Paper presented for presentation at the annual meeting of the American Educational Research Association, Toronto, Ontario, Canada, April, 1978.

A systems approach to item production and review in a computer-managed instruction project. In H. E. Mitzel (Chair), Mobile education for nurses: Computer-based instruction in support of an extended degree program for registered nurses. Symposium presented at the annual meeting of the American Educational Research Association, San Francisco, CA, April, 1976, ERIC Document No. ED 121-280.

Prayer, biographical background and college experience. Paper presented at the annual meeting of the Southeastern Psychological Association, Hollywood, FL, May, 1974.

Models for the teaching of graduate-level statistics courses in psychology departments. In E. J. Robinson, Discussion of the role of the statistics course in psychology. Symposium presented at the annual meeting of the Southeastern Psychological Association, Hollywood, FL, May, 1974.

ADMINISTRATIVE DEVELOPMENT

University of Indiana/Purdue University, The Fund Raising School, Center on Philanthropy. *Principles and Techniques of Fund Raising*. Houston, TX (March, 2005).
Lilly Foundation, *Building and supporting diversity at church-related colleges and universities*, Seguin, TX, (March, 2004).
Council of Independent Colleges and American Association of Academic Libraries, *Reforming the academic library*, San Francisco, CA, (February, 2004).
Association for Institutional Research & Council of Independent Colleges, *Data and Decisions*, Denver, CO, (September, 2003).
Council on Independent Colleges, Academic Vice President program entitled *Leading from Within*, led by Dr. Parker Palmer, Kalamazoo, MI (June, 1998).
Harvard University, *Institute for Educational Management* (July, 1995)
University of Massachusetts, Boston (New England Research Center for Higher Education), *Defining the Collective Task* (March, 1994)
Council for Advancement and Support of Education (CASE), *Major Gift Fund Raising for Deans* (May, 1993)
American Council on Education, Center for Leadership Development, *Workshop for Department and Division Chairpersons and Deans* (January, 1987)
Council of Graduate Departments of Psychology, *Workshops for Department Chairs* (February, 1986, 1987)

ADMINISTRATIVE ACCOMPLISHMENTS

Buros Center for Testing

- Brought out the 17th and 18th Mental Measurements Yearbook on time.
- Organized and ran the first strategic planning in the center's history to generate a strategic plan (2010).
- Developed a new, international vision for the Center.
- Published *Pruebas Publicadas en Español: An Index of Spanish Tests in Print*, the first such document in existence.
- Developed plans for a new institute related to assessment literacy and effected them.
- Developed plans to publish the first ever publication enumerating tests in Spanish.
- Brought in new clients for assessment outreach and consultation.
- Set a new course for the types of psychometric consultation that is appropriate given the changes in statewide testing under the Common Core.
- Developed new policies for dealing with test publishers.
- Performed outreach efforts to work with test publishers in a more effective manner.
- Hired new directors for institutes within the Center.
- Developed strategies to identify new test reviewers.
- Reorganized the meeting schedule of the National Advisory Council to provide additional external input.
- Converted secretarial position to student workers.
- Brought more than two million dollars of income over expenses over the initial four-year period.
- Began a process of succession planning.

- Updated the data base that keeps the Test Reviews and Information information.
- Regularized the meeting schedule of our National Advisory Council.
- Helped graduate assistants receive nationally prestigious summer fellowships.
- Initiated, organized and ran the first celebration of the Center's history.
- Re-oriented the Buros Institute for Assessment Consultation and Outreach to focus more on equating and validation.
- Filed suit against Taylor and Francis and received ownership of the journal, **Applied Measurement in Education** and more than tripled our income from this work.
- Developed a leadership team in the Center.

The University of St. Thomas

- Drafted, proposed and had approved a 5-year plan for the Doherty Library.
- Served as a primary participant in the SACS reaccreditation process that was completely successful.
- Partnered with the Museum of Fine Arts to provide all of our students' free membership.
- Negotiated approval of the controversial minor academic program, Woman, Culture and Society.
- Extended partnerships with other Houston museums to extend membership to students.
- Developed and instituted plan to infuse clerical support for academic units.
- Led effort with deans to develop policy on the more effective use of adjuncts.
- Coordinated and organized Chairs Workshop, Fall, 2003.
- Developed and instituted a plan to reduce Arts and Science faculty teaching loads.
- Served as member of the Governing Council, Partnership for Quality Education.
- Hired deans for Schools of Business and Theology.
- Renegotiated contract with the Diocese of Galveston/Houston for the campus of the School of Theology.
- Proposed new plan for faculty evaluation working collaboratively with the Faculty Senate.
- Radically increased grantsmanship in Academic Affairs.
- Advocated for faculty awards in the areas of teaching, scholarship and service.
- Established a process whereby goals for the Core Curriculum could be identified and agreed upon, to lead to the evaluation of the core curriculum.
- Chaired an ad-hoc committee that developed a new plan for hosting Political Speakers on campus.
- Provided Faculty Study Day (Convocation) address (Fall, 2001). Coordinated Faculty Study Day each semester.
- Held monthly open-houses for faculty members.
- Conducted focused deans' retreats each semester.
- Developed and had approved a faculty exchange program with St. Thomas University, New Brunswick, CA.
- Chaired task force on Academic Integrity. Produced recommendations for change at the University.

- Reworked the budget of the Reagan Summer Academy so that we could continue to provide collegiate instruction to underserved students from a predominantly Hispanic, urban high school.

Le Moyne College

- Initiated the O'Brien Faculty Service Award to accompany the Teacher of the Year and Scholar of the Year Awards.
- Upgraded Faculty Convocations as a true "coming together" through use of nationally recognized speakers and coordinated workshops.
- Worked with the Budget Committee and the Vice President for Finance and Treasurer to set aside funds for academic equipment (my initiative). The budget became the first academic equipment budget at Le Moyne.
- Chaired the board of a multi-university consortium (the Syracuse Consortium for the Cultural Foundations of Medicine).
- Developed and implemented a faculty-run assessment plan.
- Developed plan for, moved through governance, and initiated the Honors House across the street from the Campus Center.
- Developed and submitted a strategic plan for the Academic division. Began the development of an academic plan for the College.
- Developed a model to help identify the need for faculty lines in academic departments.
- Infused significant technology into the curriculum through a faculty development program led by a faculty member.
- Proposed and negotiated Academic Librarian Status, and had approved by the librarians, the Faculty Senate and the Board of Trustees.
- Engaged in a re-organization of the Academic Affairs Office that led to the new Assistant Vice President for Multicultural Affairs position, a return to a Director of Continuing Education position, and elimination of the Special Assistant position. Served on a committee to re-conceptualize the College into (1) Arts and Sciences and (2) Management and Graduate Studies. Hired an Associate Dean to facilitate student-centeredness.
- Worked with other administrators to increase the student-faculty ratio from 12.5-1 to 15-1, as called for by the Board of Trustees.
- Brought about and/or enhanced on-going discussions regarding new majors in Communications, Environmental Science, Global Business, Management Information Systems, Theatre, Nursing, and a masters in Accounting.
- Advanced discussion on campus concerning the arts, internships, international education, increased diversity in faculty hiring, and faculty service through my convocation speeches.
- Worked closely with a faculty committee and with input for the Trustees' Academic Affairs Committee to lead to a plan of action that led to the saving of the Physics major program.
- Led discussion that led to plans for instructional use of an older cafeteria.
- Hired a new and outstanding Director of the Madden Center (a business outreach center) from the local business community.
- Initiated accreditation effort of our Education programs through TEAC. Led discussions culminating in the decision to pursue TEAC accreditation. Moved the Department of Education substantially ahead on several fronts through the hiring of a new, outside Department Chairperson.
- Helped to fashion the Arab Studies program and to get it funded and running.

- Worked collaboratively with others to design a new Performing Arts Center and a campus archives.
- Participated in discussions related to the development of approximately 4 smart classrooms/year.
- Established annual budget line (\$150,000) for academic (e.g., scientific and instructional) equipment, not including computers (which are funded from other accounts).
- Served on a variety of American Red Cross and regional educational boards.

SUNY-Oswego

- Led conversion from a Division of Arts and Sciences to a College of Arts and Sciences through faculty governance and administrative structures.
- Raised initial funding for a speaker series to commemorate the founding of the College of Arts and Sciences. This series was so successful that it was instituted permanently.
- Increased faculty diversity significantly through hiring procedures. The percentages of ethnic minority and women tenure-track faculty hires during the 1992-93, 1993-94 and 1994-95 academic years were approximately 25% and 50%, respectively. Increased the number of women chairpersons from 0 to 3 out of 19.
- Initiated a Faculty Executive Committee for the College of Arts and Sciences to improve consultation.
- Led efforts to develop new majors in Journalism, Graphic Arts, Criminal Justice (transformed from Public Justice), Human Services, Legal Studies, Human Development, and Language and International Trade.
- Initiated and coordinated the actions leading to the chartering of Phi Kappa Phi (national honor society) and the first national interdisciplinary honor society at Oswego on campus.
- Raised non-state funds to set up a College of Arts and Sciences Faculty Development Travel Fund.
- Initiated and conducted annual new chairperson training programs.
- Organized a development program for all chairpersons from five campuses in the SUNY system. This program was evaluated by participants as being extremely effective.
- Completed development of a major in Environmental Science.
- Established a board of local health professionals to provide guidance to the health professions and to faculty in the sciences.
- Worked with the Graduate School Dean at SUNY-Health Sciences Center to initiate a summer research and pre-graduate study program for advanced students in biology, chemistry, physics and psychology.
- Co-chaired the Teacher Education Commission (1992-93).
- Hired new directors for the Tyler Art Gallery and the Rice Creek Field Station.
- Instituted more balance among teaching, service and scholarly activity through hiring policies, faculty development activities, and promotion and retention practices.
- Worked as part of a team of deans to develop a more flexible faculty workload policy.
- Initiated the effort to bring a NASA/JOVE to SUNY-Oswego and served as the administrative liaison on the JOVE team. This grant is the first NASA/JOVE grant in SUNY.
- Served on the Interim Provost Search Committee (1993-94).
- Developed a proposal, received funding for, and initiated a multi-media Language/Journalism laboratory.

- Increased the numbers of College of Arts and Sciences students studying abroad through efforts with the Office of International Education.
- Substantially updated technology within department office and science laboratories.
- Served on numerous charitable and community boards.

**UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA**

AMERICAN EDUCATIONAL RESEARCH
ASSOCIATION, INC., AMERICAN
PSYCHOLOGICAL ASSOCIATION, INC., and
NATIONAL COUNCIL ON MEASUREMENT IN
EDUCATION, INC.,

Plaintiffs,

v.

PUBLIC.RESOURCE.ORG,

Defendant.

Case No. 1:14-CV-00857-TSC-DAR

**DEFENDANT-COUNTERCLAIMANT
PUBLIC.RESOURCE.ORG'S MOTION
TO STRIKE ECF NO. 60-88, THE
DECLARATION OF KURT P.
GEISINGER IN SUPPORT OF
PLAINTIFFS' MOTION FOR
SUMMARY JUDGMENT AND
PERMANENT INJUNCTION**

Action Filed: May 23, 2014

Defendant-Counterclaimant Public.Resource.Org, Inc. ("Public Resource") respectfully moves to strike ECF No. 60-88, the Declaration of Kurt P. Geisinger In Support of Plaintiffs' Motion for Summary Judgment and Permanent Injunction.

As described in the attached Memorandum of Law in Support of Defendant's Motion to Strike, Kurt P. Geisinger's testimony includes new opinions, reasons, and facts that were not disclosed in his expert report and must be excluded under [Federal Rule of Civil Procedure 37](#). Further, Geisinger is not qualified to testify on the matters contained in the report under the standards of Federal Rule of Evidence 702 and *Daubert*. Mr. Geisinger's opinions further rest uncritically on statements from Plaintiffs' agents, invade the province of the court, and rest on unsupported assumptions, facts, and methods. For these reasons, Mr. Geisinger's report should be stricken from the record, along with all citations to and quotations of that report in Plaintiffs' Motion for Summary Judgment and Permanent Injunction.

Public Resource requests an oral hearing on this motion.

This motion is based on the enclosed Memorandum of Points & Authorities, the Declaration of Matthew Becker and the exhibits attached thereto, Public Resource's proposed Order, the pleadings and papers on file herein, and any further material and argument presented to the Court at the time of the hearing.

Dated: January 21, 2016

Respectfully submitted,

/s/ Andrew P. Bridges

Andrew P. Bridges (admitted)
abridges@fenwick.com
Sebastian E. Kaplan (*pro hac vice* pending)
skaplan@fenwick.com
Matthew Becker (admitted)
mbecker@fenwick.com
FENWICK & WEST LLP
555 California Street, 12th Floor
San Francisco, CA 94104
Telephone: (415) 875-2300
Facsimile: (415) 281-1350

Corynne McSherry (admitted *pro hac vice*)
corynne@eff.org
Mitchell L. Stoltz (D.C. Bar No. 978149)
mitch@eff.org
ELECTRONIC FRONTIER FOUNDATION
815 Eddy Street
San Francisco, CA 94109
Telephone: (415) 436-9333
Facsimile: (415) 436-9993

David Halperin (D.C. Bar No. 426078)
davidhalperindc@gmail.com
1530 P Street NW
Washington, DC 20005
Telephone: (202) 905-3434

Attorneys for Defendant-Counterclaimant
Public.Resource.Org, Inc.

MATERIAL UNDER SEAL DELETED

JA2746-JA2803

EXHIBIT 3

IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA

AMERICAN EDUCATIONAL RESEARCH)	
ASSOCIATION, INC., AMERICAN)	Civil Action No. 1:14-cv-00857-TSC-DAR
PSYCHOLOGICAL ASSOCIATION, INC.,)	
and NATIONAL COUNCIL ON)	
MEASUREMENT IN EDUCATION, INC.,)	EXPERT’S DECLARATION AND
)	REPORT OF KURT F. GEISINGER,
Plaintiffs/Counterclaim Defendants,)	Ph. D. PURSUANT TO FED. R. CIV. P.
)	26(a)(2)(B)
v.)	
)	
PUBLIC.RESOURCE.ORG, INC.,)	
)	
Defendant/Counterclaimant.)	
)	

I, KURT F. GEISINGER, Ph. D., declare:

1. I am currently Director of the Buros Center on Testing and W. C. Meierhenry Distinguished University Professor at the University of Nebraska-Lincoln.
2. The following constitutes my expert’s report in this action on behalf of Plaintiffs, the American Educational Research Association, Inc. (“AERA”), the American Psychological Association, Inc. (“APA”) and the National Council on Measurement in Education, Inc. (“NCME”) (collectively, “Plaintiffs”), complaining of certain activities engaged in by Defendant, Public.Resource.Org, Inc. (“Public Resource”).
3. This Declaration and Report contains my opinions to date. The basis for my opinions, the materials I considered in reaching my opinions, and my qualifications for rendering such opinions are set forth in this Declaration and attached Exhibits. I reserve the right to supplement my Declaration to address any additional documents and testimony introduced in this action that come to my attention between now and the time of any deposition, hearing or trial.

My Qualifications

4. I received my doctoral degree in Educational Psychology in 1977 from the Pennsylvania State University, after previously receiving my masters' degree in Psychology at the University of Georgia and my bachelor's degree from Davidson College (with honors). I also studied German, Psychology and other topics as an undergraduate at the Phillips Universität in Marburg, Germany and at Harvard University when I attended the Institute for Educational Management in 1995.

5. Previously, I served as the Vice President of Academic Affairs and Professor of Psychology at the University of St. Thomas in Houston, Texas, where I was responsible for four academic schools, approximately 200 faculty members, and over 4,000 students. I also served as Academic Vice President and Professor of Psychology at Le Moyne College, Dean of the College of Arts and Sciences and Professor of Psychology at the State University of New York at Oswego, and Professor of Psychology at Fordham University in New York City, where I was department chair for the Department of Psychology and director of the Doctoral program in Psychometrics.

6. Over the past forty years, I have researched, studied, and taught psychometrics. Psychometrics, defined in more detail later in this report, is the quantitative study of tests and measures in terms of the value, usefulness, and interpretation of the results of such measures. I also am a fellow, diplomate, and member of numerous professional societies involving educational and psychological testing, such as the APA (fellow), the American Association for Assessment Psychology (diplomate), the AERA (fellow), and the NCME, as well as other professional associations. I have represented the APA by serving on and chairing the Joint Committee on Testing Practices (which is separate from the joint committee of the AERA, the

APA and the NCME responsible for the 1999 *Standards for Educational and Psychological Testing*) and have served on the APA's Committee on Psychological Tests and Assessment. In 2010, I was elected to serve two terms (2006-2008 and 2009-2011) as the representative on the Council of Representatives for the APA's Division of Evaluation, Measurement and Statistics. My second term was cut short by one year when I was elected to serve as a member-at-large on the APA's Board of Directors in 2010, a position I held for a three-year term (2011-2013).

7. I have authored numerous publications about psychological and educational testing. I have worked at the Educational Testing Service ("ETS"), chaired its Technical Advisory Committee for the Graduate Record Examination ("GRE"), served on the Board of Directors for the GRE (a Board that I also chaired), and have been a member of the College Board, (formerly known as the College Entrance Examination Board) for which I served on its SAT Committee (from 2000-2002). I recently concluded a four-year term (from 2011-2014) on the Advisory Research Committee for the College Board, serving the last two years as its chair. I currently serve on the Technical Advisory Committee for the Educational Records Bureau.¹

8. In 2010, I was elected to the Council (i.e., Board of Directors) for the International Test Commission—the primary international testing body. In 2012, I also was elected as its Treasurer and to serve on its Executive Council. I am the only American on its Executive Council.

9. I was asked to review and share my comments on chapters of the 1999 *Standards for Educational and Psychological Testing*, published jointly by the AERA, the APA, and the

¹ The Educational Record Bureau specializes in the development and use of tests and testing products for private and independent educational institutions at the p-12 levels.

NCME (the “1999 Standards”). The *Joint Standards*² embody the professionally accepted practices for testing and measurement. One of the chapters I reviewed was based upon the testing of individuals with disabilities, an area in which I have engaged in research and have served as an expert witness in federal courts as well as state courts in New York, New Jersey, and California. The other chapter related to the rights and responsibilities of test takers. See Exh. A. I note that the *Joint Standards* were revised in 2014.

10. In addition to my 130 plus journal articles and book chapters, I have written, edited, or co-edited approximately 15 books and monographs. The vast majority of these publications deal with testing and measurement issues. For example, I have edited two books on the psychological testing of Hispanics and another I co-edited related to fairness in testing. I also have co-edited several books of reviews of published tests and measures. I also was Editor-in-Chief for the three-volume *Handbook of Testing and Assessment in Psychology* (published by the APA in 2013). Additionally, I have been editor of the journal *Applied Measurement in Education* for the past 8 plus years. Taylor & Francis, in conjunction with the Buros Center for Testing, publishes this journal.

11. I also co-chaired a sub-committee of the APA’s Joint Committee on Testing Practices and the overall committee itself that developed a document on the rights and responsibilities of test takers (from 1993-2001). This document has been endorsed by a number of professional associations related to proper test use, including the APA, the National Association of School Psychologists, the American Counseling Association, and the NCME. While chairing the Joint Committee on Testing Practices, the committee developed a book entitled *Assessing Individuals with Disabilities*, in which I wrote a chapter. I also served on a

² I use the term *Joint Standards* to refer to the *Standards for Educational and Psychological Testing* as a whole, not a specific version of the Standards, i.e. 1999 or 2014

task force charged to illuminate issues related to the testing of individuals with disabilities as well as ethnic minorities. The task force wrote and edited a book entitled *Test Interpretation and Diversity: Achieving Equity in Assessment*, which was published by the APA's publication unit in 1997. I authored three chapters in that volume.

12. I additionally served on an APA task force (from 2007-2010) that considered the assessment and intervention of individuals with disabilities. The results of our work, Guidelines for the "Assessment of and Intervention with Individuals with Disabilities," was published in the *American Psychologist*, the premier publication of the APA (Geisinger et al., 2012) and endorsed as the policy of the APA by its governance. A reference for the *American Psychologist* article may be found on my curriculum vitae, which is attached as Exhibit A.

13. In the past two years (2014-2015), I have served on two task forces related to the use of measures in clinical psychology. One of these has written a policy, recently accepted by the APA's Board of Directors, that differentiates the use of tests and other measures, for screening and assessment, two highly related types of testing, but which differ in specificity and focus. Tests are usually standardized measures that are given to a number of people for a specific purpose. A bar examination would be an example of a test. Measures are other typically quantitative values used to evaluate a person and include tests. A bathroom scale results in a measure (weight), but would not normally be considered as a test.

14. During 2013-2014, I served on a committee of the Institute of Medicine (a component of the National Academy of Sciences) that evaluated the use of psychological and clinical neuropsychological measures by the Social Security Administration in determining disability status. The final report, entitled *Psychological Testing in the Service of Disability*

Determination, is in the process of being published, but is also available from the Institute of Medicine's website.

15. For approximately four years (from 2008-2012), I jointly represented three professional associations (the AERA, the APA, and the NCME) in developing the International Organization for Standardization's ("ISO") first standard on psychological testing. The results of the work of the committee that engaged in this activity was ISO Standard 10677. The standard is divided into two parts. The first part establishes requirements and guidance for a client working with a service provider to carry out the assessment of an individual, a group, or an organization for work-related purposes. ISO 10667-1:2011 enables the client to base its decisions on sound assessment results. ISO 10667-1:2011 also specifies the responsibilities of a service provided in terms of the assessment methods and procedures that can be carried out for various work-related purposes made by or affecting individuals, groups or organizations. The second part lays out the responsibilities of the service provider in terms of the same assessment project.

16. I also developed or helped to develop a number of testing measures. Specifically, I served as the primary consultant on a number of civil service examinations given in New York City for police officer, sergeant, lieutenant, and captain, fire fighter, fire lieutenant, fire captain, sanitation supervisor, and a variety of other civil service occupations over a period of at least a decade ending in 1992. I sometimes defended these measures in court. I also represented the Public Service Alliance of Canada against the Public Service of Canada in two cases related to their national testing efforts and Disability Rights Advocates with regard to several testing disputes concerning individuals with disabilities. *See* Exh. A.

17. In recent years, my primary efforts have been to assure testing fairness for those with disabilities, language minorities, and ethnic minorities.

18. My curriculum vitae is attached to this Declaration and Report as part of Exhibit A.

19. A list of all publications that I have authored in the past 10 years is included in my curriculum vitae. *See* Exh. A.

20. To the best of my memory, during the past 4 years, I have not testified as an expert at trial or by deposition. Previously, I have been accepted as an expert on testing in state courts in New York, New Jersey, and California, and in federal courts in New York, New Jersey, and Canada. Within the past four years (2011-2015), I have been identified as an expert in cases that were settled prior to trial. I wrote a report on the use of testing to deny an individual with disability benefits for a state agency in Nebraska this past fall, but the matter was resolved prior to going to court or arbitration. In none of these cases have I given deposition testimony.

21. To explain what psychometrics is, I provide below the first two paragraphs of my entry in the *Corsini Encyclopedia of Psychology* (2010, Wiley, 3rd edition) on the topic of “Psychometrics: Norms, Reliability, Validity, and Item Analysis”.

The field of psychometrics generally considers the data from educational and psychological tests and assessments from a quantitative perspective. Such data normally emerges from test responses, although it may come from a wide variety of measurement instruments. Two divisions might be identified within psychometrics: theoretical and applied psychometrics. Psychometric theory (as portrayed by Embretson & Reise, 2000; Lord, 1980; McDonald, 1999; Nunnally, 1978) provides researchers and psychologists with mathematical models used in considering responses to individual test items, entire

tests and sets of tests. Applied psychometrics is the implementation of these models and their analytic procedures to test data (e.g., Thorndike, 1982).

22. For over five years (1989-1995), I taught courses for the Cornell University School of Industrial and Labor Relations on the topic of affirmative action and equal opportunity hiring. These courses related to the use of tests in a fair and valid way to make personnel decisions such as hiring and promotion while attempting to increase diversity and to be in conformance with federal laws and guidelines.

23. I have had past and ongoing relationships, as a member or fellow, with each of the three Plaintiff associations. I currently serve as a committee chair of one of AERA's divisions' (Division D – Measurement and Research Methodology) International Committee. I have presented at AERA's annual conferences regularly.

24. I have served on and chaired NCME's professional development committee (1990-1992), served as a program co-chair for its annual meeting (1993), ran for its board (and was defeated) (1993), represented it on the committee that developed its code of professional conduct (ethics), and was a representative and advisory board member of a doctoral program in psychometrics that was being developed at Morgan State University (2007-2012). I have published in several of NCME's journals and have served on the editorial committee for the journal, *Educational Measurement: Issues and Practice* (1992-1995).

25. I also was elected, and have served, as a member of APA's Committee on Psychological Tests and Assessment (1998-2000); on its Committee on International Relations in Psychology (2010); on its Joint Committee on Testing Practices (1992-1996), on its Council of Representatives (two terms from (2006-2010) representing the Division of Measurement, Evaluation and Statistics; and on its Board of Directors. I was appointed to serve on APA's

Good Governance Task Force that prepared a plan to reorganize its governance. I also was appointed to serve on perhaps a half dozen APA task forces over the years related to testing issues of one type or another (e.g., the testing of individuals with disabilities, the testing of individuals who are ethnic minorities, the use of testing in clinical psychology). I served for eight years (from 1992-2000) as a member of APA's editorial board for its journal, *Psychological Assessment*, and recently served on the committee that selected a new editor for that journal. Further, I served all three of these organizations by representing them on an American National Standards Institute ("ANSI")/International Organization for Standardization ("ISO") committee that developed an international standard for industrial testing.

Materials that I have Considered

26. A list of the facts, data, and materials that I have considered in forming my opinions in this case is attached to this Declaration and Report as Exhibit B.

My Opinions Relevant to this Case, and the Basis and Reasons for my Opinions

27. The *Joint Standards* serve as the foundation for the testing profession. It is the most authoritative single source about the best practices in testing. Other associations (i.e., the International Test Commission and the American Counseling Association) have much shorter and less comprehensive guidelines or standards related to testing or certain aspects of testing, but none have achieved the prominence that the *Joint Standards* currently enjoy. Also, none have the pervasive influence across different aspects of the use of tests. That is, the *Joint Standards* are appropriate in a wide range of diverse clinical, counseling, educational, and industrial settings with a variety of populations. It is because of the widespread respect in which the *Joint Standards* are held that they are sometimes cited in court cases. Elaborations of this conclusive statement follow.

28. I currently direct and have directed for the past nine years the Buros Center for Testing, formerly known as the Buros Institute of Mental Measurements. It was founded some 80 years ago by Oscar Buros, then a faculty member at the Rutgers University, to be essentially the *Consumer Reports* of the testing industry. The Buros Center publishes comprehensive, critical reviews of testing. These reviews are available through our published volumes entitled *Mental Measurement Yearbooks*. I have spoken with the other editors of our primary document, the *Mental Measurements Yearbook*, and we agree that the most commonly cited document in the reviews of tests is the *Joint Standards*. Those who review tests and testing practices refer to the *Joint Standards* constantly, and this is reflected in our *Mental Measurements Yearbooks*. We believe it to be the most frequently used comprehensive yardstick against which the quality of tests and measures and the quality of test use is evaluated.

29. In the 1990s, I co-chaired APA's Joint Committee on Testing Practices (which had no direct or formal relationship with the joint committee of the AERA, APA and NCME that develops and revises the *Joint Standards*). The role of this committee on testing practices was to develop documents and products that could improve testing practices. One document that the committee developed and subsequently revised was entitled the *Code of Fair Testing Practices in Education*, a very brief document written for parents and users of educational tests alike. The members of the Joint Committee on Testing Practices agreed that the principles espoused in the *Code of Fair Testing Practices in Education* had to be consistent with the *Joint Standards* (given their pre-eminent status). I also co-chaired a working group of that committee, which developed a document entitled *the Rights and Responsibilities of Test Takers*. We began our work this document by reviewing everything that the *Joint Standards* had to say about this aspect of testing.

30. In the multiple editions of the *Joint Standards*, various psychometric concepts have evolved or changed. Most testing experts believe that the single most important quality in testing is the validity of test scores—that they are used and interpreted in appropriate and useful ways. One can trace the history of our profession’s perceptions of validity and how it may be estimated and determined by studying its portrayals throughout the seven versions of the *Joint Standards* that have been published to date.

31. Background of the *Standards for Educational and Psychological Testing*, and its importance to the testing professions: The history of the *Joint Standards* is not brief. This history reflects changes in psychometric technology, our understandings of testing, and psychological/educational characteristics, societal fluctuations, technological improvements, as well as general zeitgeist differences. The first set of standards was published in 1954 by the APA and was entitled, *Technical Recommendations for Psychological Tests and Diagnostic Techniques*. Its impact on the testing field was monumental. Shortly after the publication of this volume in 1955, AERA and NCMUE (the original name of NCME was the National Council on Measurements Used in Education) published a similar document devoted almost exclusively to educational measures entitled, *Technical Recommendations for Achievement Tests*. There had been collaboration across these organizations in the development of these two initial documents (Eignor, 2013). The three organizations subsequently decided that their work should continue collaboratively. The two preceding documents and their first joint effort all described what test publishers should include in their test-related documentation and, to generate such information, what research efforts should be made in test development and use.

32. In 1963, the Joint Committee (across the three associations – AERA, APA and NCME) was formed, and the first *Joint Standards* were ultimately published in 1966 as the

Standards for Educational and Psychological Tests and Manuals (the “1966 Standards”). The next effort began only five years after the publication of the 1966 Standards. The Joint Committee worked from 1971 through the publication of their revised Standards in 1974 (the “1974 Standards”). This 1974 set of test standards focused less on documentation and more on topics such as how tests should best be developed, used, scored, and results reported. To emphasize the reduced focus on documentation, the title of the 1974 Standards was changed to *Standards for Educational and Psychological Tests*.

33. In the early 1980s, still another Joint Committee was empaneled and charged with the revision of the *Joint Standards*, a process that concluded with the publication of the 1985 *Standards for Educational and Psychological Testing* (“the 1985 Standards”). This minor change in the title from “tests” to “testing” emphasizes the changed focus from the tests themselves to the process of, use of, and interpretation of tests and the results of testing.

34. The 1999 Standards in question in this case held with the same title. The development process of the 1999 Standards began with open meetings where many individuals were able to speak to the Joint Committee to provide input into the ways that they believed the *Joint Standards* should change. I was one such participant at meetings in Alexandria, VA in October of 1994. Whereas the time needed to write the Test Standards had been approximately 3 years prior to the 1999 Standards, it appears that the time was closer to 4 years before the 1999 Standards became available.

35. The revision process for the current Standards that were published in 2014 was longer, although the final publication was delayed due to the present dispute with Public Resource. The Joint Committee was formalized in or around 2007, the first meeting was held in

January 2008, and the Committee completed its work in 2013. This revision process took five years.

36. The members of each Joint Committee are all volunteers, but staff support is needed. Not counting the very real costs of contributed staff support, the budget for the Joint Committee's meetings was approximately \$400,000. Had the staff salaries been covered in this budget, and had members of the Joint Committee received even minimal compensation for the work they performed, the budget for revising and updating the *Joint Standards* would have been approximately \$2,000,000. These costs will continue to increase.

37. There are several reasons I expect such cost increases. Testing is becoming both increasingly complex and increasingly technological. For example, 20 years ago all testing was "in person" or "paper-and-pencil." Now there is testing via computer, testing via tablet, and testing via phone, but the "in-person" and "paper-and-pencil" testing method still continues. Unproctored testing on the internet is presently the most common type of personnel selection testing in non-public settings in the United States. In northern Europe, the most common type of testing is now via the internet. Secondly, society is putting increasing emphasis upon measurement concerns. Teachers are being evaluated in many settings based upon how their students perform on tests. Every year more professions and positions require tests to justify access to positions (e.g., via licensure and certification testing). As more and more testing cases are litigated, the need for clarification of professional practice increases concomitantly. Finally, travel and hotel expenses only continue to rise.

38. Prior to the 1985 Standards, all of the individual standards composing the Standards volume were considered separately as "essential, very desirable, or desirable". These statements indicated relative degrees of importance. The 1985 Standards were identified as

primary, secondary, or conditional, depending upon both the importance and the nature of their use. Beginning in 1999, no status was assigned to individual standards, a practice that was also continued for the 2014 Standards. I heard informally from members of the 1999 Joint Committee that the reason for the change was to de-emphasize the role that the *Joint Standards* might play in litigation.

39. One can glean from the prior discussion that the *Joint Standards* are changed and updated approximately every 10-15 years. The revision process has been taking an ever-longer amount of time, probably due at least in part to the increased focus on educational testing in the accountability movement across education in the United States. Testing is now used in part to affect federal budgets allocated to states to provide education and, for example, for teacher and professional staff evaluation. It is likely that the time frame will continue to increase as the focus on educational testing shows no letup in the foreseeable future. In fact, it is clear that the *Joint Standards* were once revised every 10 years and for the past two revisions, it has taken about 15 years. One reason for this temporal increase relates to the complexity of the changes.

40. Having three important professional associations involved in the development and updating of the *Joint Standards* helps bring much credence to them. It also, however, makes the process more cumbersome due to the communications and decision-making necessitated by having three professional associations involved that may all see various professional issues differently.

41. Since the revision of the 1999 Standards, the development of revised *Joint Standards* is controlled by a Management Committee. The three associations (AERA, APA, and NCME) established a Management Committee consisting of three representatives, one from each sponsoring organization. These members are appointed by the Chief Executive Officers of

AERA and APA and by the President of NCME. Each member normally serves a 3-year, renewable term. Members are usually appointed so that terms are staggered, providing for continuity. Members may be reappointed for one additional three-year term, and in years when the *Joint Standards* are in active revision, members' terms for the duration of the publication period may be extended by each sponsoring organization.

42. Members of the Management Committee oversee all aspects of the *Joint Standards* on an ongoing basis, including, but not limited to: publication and distribution, oversight of the Development Fund, protecting the copyright of the Standards, archival activities, gathering information about their use, and potential revision issues. The Management Committee represents both the interests of the *Joint Standards* and the interests of all three sponsoring organizations in conducting its administrative duties.

43. At least once every five years, members of the Management Committee confer with the leadership of their respective organizations to assess the need for revision of the *Joint Standards*.

44. If it is determined that the *Joint Standards* need revision, the Management Committee first appoints co-chairs of a "Joint Committee" that addresses the specifics of the revision effort. In the case of the 2014 Standards, a website was developed whereby members of the sponsoring organizations could provide feedback concerning the changes that they believe necessary or important. Following this process, in collaboration with the co-chairs, the Management Committee appoints members of the Joint Committee who represent the sponsoring organizations. The Joint Committee provides a structure for communication with the sponsoring organizations throughout the revision process.

45. One of the reasons that the process is so lengthy is that the Joint Committee is composed of nationally prominent experts cutting across clinical psychology, counseling psychology, school psychology, industrial/organizational psychology, clinical neuropsychology, and educational testing. The Joint Committee is composed of highly regarded professionals at the top of their respective fields, who have specialized knowledge regarding the use of tests and measures in their respective disciplines. Getting such extraordinarily busy individuals to agree to serve in a volunteer, unpaid fashion is difficult enough. Working with their schedules to set up meetings when all can attend is administratively an arduous process. My understanding is that the individuals take on this task as service to their profession, their associations, and the *Joint Standards* themselves.

46. The Management Committee oversees the process throughout the development of new standards, and reports back to their respective associations.

47. The entire *Joint Standards* revision process is financed through sales of a prior version of the *Joint Standards*. As noted above, the direct costs for the development process has been estimated at \$400,000 for the 2014 Standards and between \$500,000 and \$600,000 for the 1999 Standards.

48. What Public Resource did with the 1999 Standards: I reviewed Public Resource's discovery responses and the transcripts (with exhibits) from the depositions of Carl Malamud (President and Founder of Public Resource) and Chris Butler (Office Manager of the Internet Archive). In May 2012, I understand that Mr. Malamud purchased a used copy of the 1999 Standards. I understand further that he then sliced the printed pages out from their bindings, scanned the pages to a PDF file, and posted the file to Public Resource's website as well as to a publicly available collection on the Internet Archive. See Exh. C at 5-6. A graduate psychology

student who engaged in such actions would probably be dismissed from his or her program. Most certainly he or she would be subject to ethics charges that could follow them throughout the person's career.

49. The PDF file posted to Public Resource's website and to the Internet Archive contained a cover page prepared by Mr. Malamud, leading Internet users who came upon it to believe that the 1999 Standards were freely available for download, copying, or whatever use someone wanted to make of the text. *See* Exh. D. To the extent that individuals accessed the 1999 Standards in this fashion, it would appear to be theft of services.

50. The PDF file with Mr. Malamud's cover page was posted to Public Resource's website and to the Internet Archive from July 2012 until June 2014. *See* Exh. C at 5. Based upon incomplete records provided by Public Resource, during the time it was posted to the Internet, the PDF file containing the entire text of the 1999 Standards that was posted to Public Resource's website was accessed by Internet users at least 4,405 times. *See* Exh. C at 9-10. The same PDF file containing the entire text of the 1999 Standards that was posted to the Internet Archive website was accessed by Internet users at least 1,113 times. *See* Exh. E. No restrictions were placed on this PDF file to prevent Internet users from downloading the file to their local hard drives or printing it using a printer attached to an Internet user's computer. *See* Exh. F at 346-47. These accessions represent considerable lost revenue to the organizations supporting the Joint Committee. However, even if the incomplete records kept by Public Resource and Internet Archive were completely accurate, such numbers would not represent the real numbers as lost revenue because one person can download the *Joint Standards* and share them with hundreds of colleagues.

51. In December 2013, AERA asked Mr. Malamud to remove the 1999 Standards from the Internet locations where he posted the document. He refused. Ultimately, Mr. Malamud did remove the 1999 Standards from public view on the Internet, but only after Public Resource was sued for copyright infringement and threatened with a motion for a preliminary injunction. *See* Exhs. F at 324-26, G.

52. During his deposition, Mr. Malamud testified that, should Public Resource succeed in this litigation, it would be a very easy matter for him to re-post the 1999 Standards to his company's website and to the Internet Archive website. *See* Exh. F at 307. Further, Mr. Malamud contemplated that Public Resource might do the same with the 2014 Standards. *See* Exh. F at 308-09. Such actions would almost certainly lead to the 2014 Standards being the last one developed or published. I have heard from definitive sources within the three organizations (AERA, APA and NCME) that without the revenue from the sale of the *Joint Standards*, there would not be funding from other sources to continue updating them. There simply would not be funding to continue the updates without these needed sales revenues.

53. Given the changes happening concomitantly in testing, the testing profession, and society generally, the cessation of updates to the *Joint Standards* would be a travesty. The profession relies on the *Joint Standards* increasingly in a changing world where high stakes decisions are often buttressed by information gleaned from tests and measures. Indeed, the *Joint Standards* are needed. The *Joint Standards* are critically important to professionals who work with tests and measures in education and psychology.

54. Public Resource's justification for posting the 1999 Standards to the Internet: It is Public Resource's view that, once the 1999 Standards were incorporated by reference into federal and/or state regulations, the 1999 Standards lost their copyright protection. As a

consequence, Public Resource believes that it and others can freely reproduce the 1999 Standards (in this case, in electronic format), and post the document to the Internet so that it is freely available to everyone. *See* Exh. H.

55. The past and continued harm that electronically reproducing and posting the 1999 Standards to the Internet will cause AERA, APA and NCME: The three associations that are the Plaintiffs in this case (AERA, APA, and NCME) are integrally involved in the revision and publishing of the *Joint Standards*. I have discussed the prospects of the continuation of the *Joint Standards* with knowledgeable representatives of all three organizations. I fear that the 2014 Standards will be the final version should AERA's, APA's, and NCME's copyright infringement claims against Public Resource not succeed. The loss of income caused by the document being made freely available has already had a significant and negative impact. The loss of sales revenue negatively affected the three associations' budget for the development of the 2014 Standards, and cost the associations some credibility for seeming to permit an organization such as Public Resource to violate copyrights that the testing profession considers so sacred (because tests too are copyrighted).

56. Almost certainly, none of the three associations would be willing or able to finance the continuation of revisions to the *Joint Standards* if they are made freely available. The Plaintiff associations would not be financially able to continue the re-development process into the future. None of the associations would even have the inclination of their governance or membership to carry on with publishing the *Joint Standards*, given the additional burden this would place on membership costs.

57. The past and continued harm that electronically reproducing and posting the 1999 Test Standards to the Internet will cause to the testing professions and the public: The primary

consequences of not revising the *Joint Standards* would be twofold: to the public, who are impacted by changes in testing practices, and to test users and their clients. Because of society's reliance on test results, a significant portion of the population is benefitted by proper testing practices (e.g., employers select the best and most appropriate job candidates, colleges and universities choose the applicants most likely to succeed in their programs, students receive credit for their learning, programs can assess their successfulness). Changes are necessitated to testing practices when societal norms and technology change. Recent editions of the *Joint Standards* have included chapters on the testing of ethnic minorities, individuals with disabilities, and fairness, for example. The *Joint Standards* represent something of a gold standard to which test developers and users aspire. The *Joint Standards* also have been modified as the needs for various kinds of measurement have changed. Should the practice of posting the *Joint Standards* to the Internet continue, it is likely that there will be no formally sanctioned process for their continuation. The agency that I run is essentially one for consumer protection. Without the *Joint Standards*, I fear that many customers (clinical psychologists, counseling psychologists, industrial psychologists, school psychologists, test developers, psychometricians, and the organizations that each works in) will be the ones losing.

Conclusions

58. The *Joint Standards* represent the single best and most complete statement of how tests and other measurements should be developed, used, evaluated, and interpreted. The *Joint Standards* have a long development history (for the social sciences). It is a history that is currently endangered by what I consider copyright theft. We in academe and the scholarly professions often report that all we have is our ideas and our writing. If someone is able to steal our ideas so openly and callously, our professions and indeed our society suffer.

59. AERA, APA, and NCME have engaged in a laborious and time consuming project (actually a history of projects) that they expected to be rewarded with resultant modest revenues. The vast portion of these revenues has gone to funding the process for the continual revision of the *Joint Standards*. Without such a revenue stream, the *Joint Standards* may end with the current edition.

60. If there is not a next edition of the *Joint Standards* in the 2020s, then needed changes in professional practice would not be acknowledged in as formal and yet aspirational a manner as permitted by the *Joint Standards*. While the *Joint Standards* are not published to make money *per se*, there is an expectation of modest revenues. The *Joint Standards* are debated, considered, and written to improve practice. Funding is needed to continue this effort.

61. Extremely well qualified members of our professions are willing to volunteer their time to serve on the Joint Committee to work on the *Joint Standards*. They do so for two primary reasons: i) to improve professional practice in their area of expertise, and ii) to benefit their professional associations. If the latter goal is removed due to lost revenues to the professional associations, then the quality of those willing to serve is likely to be reduced. People may only engage in this work if they are compensated, which again would be extremely difficult given the lack of or severe reduction in revenues.

62. The current *Joint Standards* were published in 2014. Yet the version that Public Resource placed on the Internet for free access was the 1999 version of the *Joint Standards*. Unsuspecting people (e.g., students) may well access these freely available standards and believe that they are the “current” *Joint Standards*. Moreover, given the high stakes nature of our society presently, suppose a small test developer (and there are many of them) accessed the outdated *Joint Standards* and used them to develop, use, and/or interpret the results of a test. It

is possible that their decisions and actions would be out of date. Further, given that the *Joint Standards* have been quoted as having been given great deference by the courts, it is possible that should a situation like the above occur, a test developer or test user could be placed in a situation whereby the advice that they follow is outdated, perhaps inappropriate advice. They could even be subjected to legal liability for such a well-intended, if somewhat naïve, action.

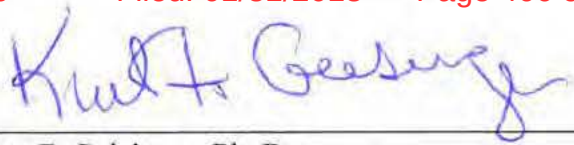
63. I expect that when one successively publishes a document and revisions to that document, one has a reasonable expectation that the publisher does so on the basis of exclusivity. No one else can publish that same document. The *Mental Measurement Yearbooks*, which the Buros Center for Testing that I direct publishes, lost considerable money in its first four or five editions. We are now publishing the 20th edition, and we have the expectation to earn revenues with each successive edition. I would hope that if someone else would be prohibited from simply copying our books and distributing them for free. The well-earned reputation of solid and professional work should be rewarded with an expectation of a fair return. Giving an item away for free, such as the *Joint Standards*, violates that very principle. If the revenue loss is serious enough, it may remove the desire and expectation for the sponsoring organizations (AERA, APA and NCME) to continue publishing the document. That is exactly the situation in which these three professional organizations find themselves.

Engagement and Compensation

64. A Letter of Agreement engaging my services in this action and stating the compensation that I will be paid for my study and testimony in the case is attached to this Report as Exhibit I.

I DECLARE, and the penalty of perjury, that the foregoing is true and correct.

Dated: June 10, 2015



Kurt F. Geisinger, Ph. D.

EXHIBIT 4

EXHIBIT B

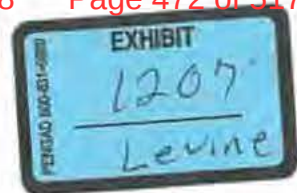
Case No. 1:14-cv-00857-TSC-DAR

List of Materials Considered

1. AERA, APA and NCME's Complaint – 5/23/14;
2. Exhibits A and B to AERA, APA and NCME's Complaint – the Copyright Registrations issued for the 1999 Standards;
3. The Counterclaim and Answer to the Complaint of Defendant, Public.Resource.Org, Inc. ("Public Resource") – 7/14/14;
4. AERA, APA and NCME's Reply to Public Resource's Counterclaim – 8/21/14;
5. AERA, APA and NCME's Amended Disclosures – 05/18/15;
6. Public Resource's Initial Disclosures – 05/18/15;
7. Public Resource's Amended Interrogatory Answers (1st Set) – 12/15/14;
8. Public Resource's Admissions' Responses – 11/3/14;
9. Public Resource's Interrogatory Answers (2nd Set) – 3/2/15;
10. Public Resource's Amended Answer to Interrogatory No. 8 – 6/4/15;
11. AERA, APA and NCME's Interrogatory Answers – 1/20/15;
12. AERA, APA and NCME's Admissions' Responses – 1/20/15;
13. The transcript and exhibits from the deposition of the Internet Archive (by Christopher Butler) taken on December 2, 2014;
14. The transcript and exhibits from the deposition of the Public Resource (by Carl Malamud) taken on May 12, 2015;
15. Conversations with Felice J. Levine, Ph. D., Executive Director of AERA. It is my understanding that AERA is the publisher of the 1999 and 2014 Standards;

16. DANIEL R. EIGNOR, *The Standards for Educational and Psychological Testing*, APA HANDBOOK OF TESTING AND ASSESSMENT IN PSYCHOLOGY, VOL. 1 at 245-250, (K. F. Geisinger ed., American Psychological Association, 2013);
17. SUSAN E. EMBRETSON & STEVEN P. REISE, PSYCHOMETRIC METHODS: ITEM RESPONSE THEORY FOR PSYCHOLOGISTS (Lawrence Erlbaum Associates, Inc. 2000);
18. FREDERIC M. LORD, APPLICATIONS OF ITEM RESPONSE THEORY TO PRACTICAL TESTING PROBLEMS (Lawrence Erlbaum Associates, Inc. 1980);
19. RODERICK P. McDONALD, TEST THEORY: A UNIFIED TREATMENT (Taylor & Francis 1999);
20. JUM C. NUNNALLY, PSYCHOMETRIC THEORY (2d ed., McGraw-Hill 1978);
21. ROBERT L. THORNDIKE, APPLIED PSYCHOMETRICS (Houghton Mifflin Co. 1982).

EXHIBIT 5



**Standards for Educational and Psychological Testing
 Sales Report, 1999 Edition**

Period	Notes	No. of Units
FY 1999	est.	1,768
FY 2000	est.	3,797
FY 2001	est.	3,755
FY 2002	est.	5,592
FY 2003	est.	3,310
FY 2004	est.	3,218
FY 2005	Actual	3,803
FY 2006	Actual	3,888
7/1/06-12/31/06	Actual	2,144
FY 2007	Actual	3,077
FY 2008	Actual	3,358
FY 2009	Actual	2,590
FY 2010	Actual	3,043
FY 2011	Actual	2,132
FY 2012	Actual	1,649
FY 2013	Actual	1,732
FY 2014	Actual	855
Total Units Sold		49,710

Note: Estimates are based on revenue earned and reported.

MATERIAL UNDER SEAL DELETED

JA2834-JA2837

EXHIBIT 8



APA Membership Statistics

Year	Associates	Members	Fellows	Total
2014	7,866	62,924	4,449	79,796
2013	8,350	69,248	4,555	82,153
2012	8,535	70,054	4,491	83,080
2011	8,593	71,247	4,499	84,339
2010	9,223	77,508	4,626	91,306
2009	8,775	78,618	4,626	92,019
2008	8,318	79,152	4,852	92,322
2007	7,943	79,407	4,705	92,055
2006	7,385	79,158	4,653	91,196
2005	7,056	78,542	4,658	90,256
2004	7,144	78,416	4,642	90,202
2003	7,240	77,938	4,597	89,775
2002	7,507	77,316	4,580	89,403
2001 ^{30, 31, 32}	7,618	76,660	4,547	88,825
2000	6,732	71,847	4,517	83,096
1999	7,068	72,064	4,484	83,617
1998	7,165	71,364	4,409	82,938
1997	7,450	70,587	4,350	82,387
1996	7,841	69,335	4,355	81,531
1995	7,719	67,063	4,316	79,098
1994	7,532	64,234	4,242	76,008
1993	7,295	61,806	4,162	73,263
1992	7,631	60,892	4,121	72,644
1991	7,884	60,259	4,059	72,202
1990	7,903	58,311	4,052	70,266
1989	8,098	56,226	3,997	68,321
1988	8,347	54,644	4,005	66,996

JA2839

1987	8,823	52,584	3,737	65,144
1986	8,587	50,727	3,832	63,146
1985	8,511	47,901	3,719	60,131
1984	8,539	46,042	3,641	58,222
1983	8,600	44,212	3,590	56,402
1982	8,681	42,071	3,528	54,282
1981	8,706	40,301	3,433	52,440
1980	8,865	38,675	3,393	50,933
1979	8,909	36,804	3,333	49,047
1978	8,817	34,832	3,242	46,891
1977	8,658	32,797	3,195	44,650
1976	8,278	30,576	3,174	42,028
1975	7,795	28,552	3,064	39,411
1974	7,357	26,644	2,999	37,000
1973	7,052	25,243	2,959	35,254
1972	6,832	23,870	2,927	33,629
1971	6,611	22,526	2,848	31,985
1970	6,532	21,502	2,805	30,839
1969	6,070	19,909	2,806	28,785
1968	5,640	18,889	2,721	27,250
1967	5,219	17,955	2,626	25,800
1966	4,812	17,095	2,566	24,473
1965	4,362	16,664	2,535	23,561
1964	3,791	15,865	2,463	22,119
1963	3,213	15,342	2,378	20,933
1962	2,623	14,931	2,337	19,891
1961	2,033	14,640	2,275	18,948
1960	1,408	14,569	2,238	18,215
1959	744	14,485	2,219	17,448
1958	none	14,474	2,170	16,644
1957	13,457		2,088	15,545
1956	12,503		2,006	14,509
1955	11,579		1,896	13,475
1954 ²⁹	10,567		1,813	12,380
1953	9,233		1,690	10,903

1952	7,927		1,585	9,512
1951	6,979		1,576	8,554
1950	5,775		1,498	7,272
1949	5,299		1,436	6,735
1948	4,493		1,261	5,754
1947	3,583		1,078	4,661
1946 ²⁸	3,344		1,083	4,427
1945	3,161	1,012		4,173
1944	2,948	858		3,806
1943	2,716	760		3,231
1942	2,518	713		3,231
1941	2,254	683		2,937
1940	2,075	664		2,739
1939 ²⁷	1,909	618		2,527
1938 ²⁶	1,715	603		2,318
1937	1,551	587		2,138
1936	1,431	556		1,987
1935	1,276	542		1,818
1934	1,224	530		1,754
1933	1,135	535		1,670
1932	985	525		1,510
1931	737	530		1,267
1930	571	530		1,101
1929	353	540		893
1928 ²⁵	165	534		699
1927 ²³	92	516		608 ²⁴
1926	41	494		535
1925 ²²		471		471
1924 ²⁰		464		464 ²¹
1923 ¹⁷		457 ¹⁸		457 ¹⁹
1922		442		442
1921 ^{14, 15}		424		424 ¹⁶
1920 ¹³		393		393
1919		372		372

1918	367	367	367
1917		336	336
1916 ¹¹		308	308 ¹²
1915 ¹⁰		291	291
1914		285	285
1913		271	271
1912		262	262
1911 ⁹		244	244
1910		228	228
1909		225	225
1908		209	209
1907		209	209
1906 ⁸		190	190
1905		168	-168
1904		94	-151
1903		135	135
1902		127	127
1901		127	127
1900		127	127
1899		113	113
1898		111	111
1897 ⁷		87	87
1896 ⁶		94	94
1895		78	78
1894 ⁵		67	67
1893		54 ³	54
1892		31 ¹	31
1892		42 ²	42 ⁴

Footnotes

¹ Preliminary Meeting.

² First Annual Meeting.

³ Figures in parentheses are estimates.

⁴ The first mention of membership appears in a tentative ad interim constitution adopted at the first annual meeting (1892) which reads: "The right of nomination for membership is reserved to the Council, the election to be made by the Association." (Fernberger, 1932, p. 7-8).

- 5 In the first Constitution adopted at the third meeting (1894) no specific article is concerned with membership. But in Article II, which provides for a council of six members with the president ex-officio, we find as one of its duties that they "shall nominate new members" and also that "the resolutions of the Council shall be brought before the Association and decided by a majority vote." (Fernberger, 1932, p. 8).
- 6 As early as 1896, one finds that (Lightner) Witmer proposed that "all names nominated by the Council, shall be presented to the Association at its opening meeting in written form or visibly displayed upon a blackboard, together with a statement of the contribution or contributions to psychology, in virtue of which the persons named are eligible to Membership, and that the final action upon such names shall be taken by the Association at the final business meeting." (Fernberger, 1932, p. 8).
- 7 Perhaps because of (Lightner) Witmer's motion the previous year, it was voted in 1897 "that nomination blanks be provided by the Secretary with spaces for the name, official position and publications of the candidate and the names of two proposers, members of the Association; such blanks to be filled in and sent to the Secretary before the meeting and to be read before the Association when the name of such candidate comes up for election." (Fernberger, 1932, p. 8).
- 8 Council decided in the future to define the qualifications and make them more difficult. This was accomplished in 1906 by a formal announcement of the council to the association of the principles which guided them in nominating or declining to nominate individuals proposed for membership. "The Constitution reads that those are eligible for membership who are engaged in 'the advancement of Psychology as Science.' In interpreting the Constitution the Council has, historically and consistently, recognized two sorts of qualifications for membership: professional occupation in psychology and research. The Council now adheres to a somewhat strict interpretation of the former of these qualifications so that, in the absence of research, positions held in related branches such as philosophy and education, or temporary positions, such as assistantships in psychology, are not regarded as qualifying candidates for membership." (Fernberger, 1932, p. 9).
- 9 "The Council having for some years back experienced frequent difficulty in securing adequate information regarding applicants for membership in the Association made public the following announcement: The Council requests that all recommendations for membership in the Association submitted to the Secretary at least one month in advance of the time of election, and that these recommendations be accompanied by Statement of the candidate's professional position and by copies of published researches." (Fernberger, 1932, p. 9).
- 10 In 1915, at the end of this low period, (Charles) Judd questioned the council's interpretation of a statement regarding requirements of candidates for admission to membership in the association and moved that it be the sense of the association that the statement appended to Article I of the Constitution defining 'temporary positions' should be interpreted to include under this head the position of instructor." The motion was carried and we see, for the first time, the association as a whole, rather than the council, initiating a definition of qualifications for membership. This motion defines an instructorship as a temporary position and hence, for a younger man, throws still greater emphasis on the question of publication. (Fernberger, 1932, p. 10).
- 11 In the next year (1916) the council again initiates a move for greater standardization as follows: "A proposal for membership, signed by at least two members of the Association, must be submitted to the Secretary, for the Council at least one month in advance of the annual meeting. The proposal must be accompanied by (1) a statement of the candidate's professional position and degrees, naming the institutions by which and the dates when, conferred, and (2) by copies of his published researches. In the absence of acceptable publications of a psychological character, or a permanent position in psychology, the conditions of membership will not be regarded as having been fulfilled." This announcement merely still further defined Judd's motion of the year before and for the first time specifically mentions academic degrees. (Fernberger, 1932, p. 10).
- 12 In the same year (1916) the council also announced that "Proposals to membership that are unfavorably acted upon by the Council must be renewed for action at a subsequent meeting." (Fernberger, 1933, p. 10).
- 13 In this year (1920) it was voted "that a committee of three, including the Secretary, be appointed by the President to revise the requirements for membership and to report at the next annual meeting of the Association." Boring was appointed chairman with Dunlap and Terman as the committee. It was also proposed and voted that this be referred to the new committee, that foreign members be not elected to active membership but "that distinguished psychologists in foreign countries be elected, upon recommendation of the Council, corresponding members of the Association and that such corresponding members be not subject to the payment of dues." (Fernberger, 1932, p. 11).
- 14 In 1921 this committee reported and the report was adopted by the association in part only. The committee recommended two grades of membership, members and fellows. The recommendation was for the creation of 100 fellows within the membership of the association and asked for a new committee to consider the mode of election of these fellows, their qualifications, functions, etc. (Fernberger, 1932, p. 11).
- 15 But the first part of the report, which was adopted and became law, more fully and clearly defines qualifications for membership.

UBCA Case #17-1055 Document #17-15850 Filed: 01/01/2016 Page 480 of 517

In a preamble the committee states "The Committee believes further that the qualifications should be formulated in accordance with the object of the Association, 'the advancement of psychology as a science' as stated in the Constitution; and they believe that this end will be most readily secured by placing emphasis upon scientific publication. They believe further that the time has come to abandon professional position or title as a basis for election on account of the reason that the multiplication of special positions, especially in nonacademic fields of psychology, makes the interpretation of the significance of position impracticable." In order to enforce this point of view, the Association adopted the Committee's specific recommendations for qualifications for members the establishment of an 'associate' grade of membership and to report to the 1924 meeting with recommendations." (Fernberger, 1932, p. 11-12).

¹⁶ The Association adopted the committee's specific recommendations for qualifications for membership which were "(1) acceptable published research of a psychological character and (2) of the degree of the Doctor of Philosophy, based in part on a psychological dissertation." The question of the degree may be waived by the council in special cases providing it states its reasons when making the nomination. And further "(3) it is also expected that the Council shall assure itself that the nominee is actively engaged in psychological work at the time of the nomination." (Fernberger, 1932, p. 12).

¹⁷ 1924: At the meeting the year before it was decided that nominations must be made "not later than March 15 of the year in which the nomination is to be first acted upon." (Fernberger, 1932, p. 12).

¹⁸ 1923: the Council shall have power to defer action upon such proposals for membership as it deems necessary providing, however, that the third annual meeting after the original receipt of the nomination papers, it must decide either to present or not to present the candidate's name to the Association. A proposal for membership cannot be reviewed until two years have elapsed after the Council's action upon it." (Fernberger, 1932, p. 12).

¹⁹ 1923: It was voted that a committee of three be appointed "to consider the advisability of the establishment of an 'associate' grade of membership and to report to the 1924 meeting with recommendations." Boring was appointed chairman of this committee with F. L. Wells and Hunter. The report, which was a lengthy one, was presented in 1924 and printed in the Proceedings. The committee "are unanimous in the opinion that the purposes of the Association will be served by the creation of a class of Associates " because the growth of psychology has "created distinct groups of persons engaged in psychological work of a scientific character at less advanced levels" so that the fundamental requirements of membership can no longer be met by this group. Hence the Committee proposes a class of Associates eligible under the following qualifications: "(1) any person devoting full time to work that is primarily psychological; (2) any person with the degree of Doctor of Philosophy, based in part upon a psychological dissertation and conferred by a graduate school of recognized standing, or (3) scientists, educators or distinguished persons, whom the Council may recommend for sufficient reason." (Fernberger, 1932, p. 12).

²⁰ The exclusionary tendency that predominated the first two decades of the 20th century was to eliminate from membership individuals who were not directly involved in psychological pursuits. The Definition of Psychology officially hinged on the terminology of the association's constitution as "The Advancement of Psychology as a Science," which was primarily that of academic psychology involved in research, primarily experimental research. In general, it was the individuals on the periphery of psychology who were eliminated, those with a non professional, amateur's interest in the field, and those primarily involved in philosophy. (Evans, 1992, p. 78).

²¹ The committee then further recommends certain methods of application of the change. The application for associateship may be made by the candidate rather than by two proposers as for membership. But two endorsers must be specified by the applicant with whom the council may (and always did) communicate. The application must be received by October 1 instead of March 15 as for Members. The council is to consider all applications for associateship and recommend to the association which elects. The associates to have the right of the floor at the annual meetings and the right participate in the programs but are not entitled to hold office or to vote. Upon the recommendation of the council and by the majority vote of the annual meeting an associateship may be terminated. (Fernberger, 1932, p. 13).

²² The necessary by-laws and constitutional changes were passed for the first time in 1924 and received the necessary second passage in 1925. Immediately and at the same meeting these changes the by-laws became effective by the election of forty-five associates. (Fernberger, 1932, p. 13).

²³ The committee suggests a form by means of which associates may apply for membership. This is to be accomplished by having all associates asked each year if they care to make application for membership. The committee also suggested a similar form of application blank for both grades. The changes were passed in 1927 on its second reading. This change had the effect of still further raising the qualifications for Membership by defining a policy of the council demanding at least two publications beyond the doctorate thesis. It makes the date of application for both grades uniform with a closing on March 15th. (Fernberger, 1932, p. 14).

²⁴ The council in 1927 were willing to recommend only a relatively few associates for membership inasmuch as they were not willing to construe graduate work as "devoting full time to professional work in psychology." Hence in this year a change was made

in the bylaws which changed this qualification to read "who have had at least one full year of graduate work in psychology in a recognized graduate school or who at the time of application are devoting full time to professional or graduate work in psychology." (Fernberger, 1932, p. 14).

²⁵ In 1928 a new mechanism for handling nominations was approved by the council. According to this new method, which is still in practice, the Secretary first reviews each nomination. For those cases where there is no question that the candidate is eligible for associateship but not for membership (and this includes the great majority of the cases) the secretary himself approves the nomination and writes to so inform the candidate, telling him that if he objects to this ruling and insists upon being considered for membership, that his case will be presented to the council. For all other cases, those who seem to be eligible for membership and those whom the secretary considers are not qualified for associateship, the former method of submitting transcripts for the consideration of the council is followed. (Fernberger, 1932, p. 15).

²⁶ 1) The association shall consist of three classes of persons: first, members, second, associates and third, honorary members. 2) Members of the association shall be persons who are primarily engaged in the advancement of psychology as a science. 3) Associates shall be such other persons as are interested in the advancement of Psychology as a science and who desire affiliation with the association for this reason. Three honorary members shall be persons, who having reached the age of seventy years and having been members for at least twenty years, request such status. (APA Yearbook, 1938, pgs. 14-15).

²⁷ The association shall consist of three classes of persons: first, members, second, associates and third, life members. Four life members shall be persons who, having reached the ages of seventy years and having been members of the association for at least twenty years, request such status. (APA Yearbook, 1939, pg. 21).

²⁸ The association shall consist of three classes of members: Fellows, associates and life members. Two fellows of the association shall be persons who are primarily engaged in the advancement of psychology as a profession..(APA Yearbook, 1946-1947, p. 26).

²⁹ 1954, the council formally requested the Policy and Planning board to study the standards for membership, which, at that time, were those set forth in article II of the original (1946) bylaws. These classes of Membership were defined as follows:

- Fellow. Holder of Doctoral degree based in part of a dissertation psychological in nature, prior membership as an associate and acceptable, published research beyond the dissertation or four years of acceptable professional experience. The nomination was made by a Division to the Board of Directors, which, if approved was recommended to the council.
- Associate. Holder of a doctorate *or* completion of two years of graduate work in psychology, *or* completion of the year of graduate study and one year of professional experience; or that the individual be a distinguished person recommended by the board of directors.
- Life Member. A fellow or and associate for 25 years and attainment at age 65.

As a result of its deliberations, the Policy and Planning board recommended to the Board of Directors that the categories be revised. After some years of debate, the Council approved three classes of membership: fellow, member and associate. On approval by the membership, this change went into effect at the beginning of 1958. Standards for Fellow were strengthened by requiring the nominating division to furnish the Membership Committee with clear evidence of the candidate's unusual or outstanding accomplishment in Psychology. The new category of member required the doctorate, thus preserving the time-honored criterion. The class of associate was continued for subdoctoral psychologist, but it was stipulated that when an associate was awarded the doctorate, he or she would automatically be raised to member. The life member category was dropped, but waiver of dues, when requested, for members over 65 years of age and with 25 years of membership were retained. Various types of affiliates, such as student, division and foreign were recognized, but, as in 1945, they were not counted as members of the association. (Evans, 1992, p. 182-183).

³⁰ Member: The minimum standard for election to member status is receipt of the doctoral degree based in part on a psychological dissertation or based on other evidence of proficiency in psychological scholarship. The doctoral degree must be received from a program primarily psychological in content and must be conferred by a graduate or professional school that (a) is regionally accredited or (b) has achieved such accreditation within five years of the year the doctorate was granted, or (c) is a school of equivalent standing outside of the United States. All members may vote and hold office in the association. (Directory, 2001, p. IX).

³¹ Associate Member: To become an associate member, an applicant must meet one of two sets of requirements: (a) must have completed two years of graduate work in psychology at a regionally accredited graduate or professional school or (b) must have received the master's degree in psychology from a regionally accredited graduate or professional school. Associate members initially may not vote or hold office in APA. After five consecutive years of membership, associate members may vote. (Directory, 2001, p. IX).

32 Fellow US affiliated members may, on nomination by an APA Division and election by the Council of Representatives, become fellows of the APA. Candidates for fellows status must previously have been members for at least one full year, have a doctoral degree in psychology and at least five years of acceptable experience beyond that degree, hold membership in the nominating division, and present evidence of unusual or outstanding contribution or performance in the field of psychology. Fellows may vote and hold office. (Directory, 2001, p. IX).

Bibliography

- American Psychological Association. (1938). "Bylaws: Article I." *American Psychological Association Yearbook: 1938 Edition*. Washington, DC: American Psychological Association. 14-15.
- American Psychological Association. (1939). "Bylaws: Article I." *American Psychological Association Yearbook: 1939 Edition*. Washington, DC: American Psychological Association. 21.
- American Psychological Association. (1947). "Bylaws: Article II." *American Psychological Association Yearbook: 1946-1947 Edition*. Washington, DC: American Psychological Association. 26-27.
- American Psychological Association. (2001). *Directory of the American Psychological Association: 2001 Edition*. Washington, DC: American Psychological Association. IX.
- Crawford, Meredith P. (1992) "Rapid Growth and Change." *100 Years: The American Psychological Association: A Historical Perspective*. Washington, DC: American Psychological Association. 182-183.
- Evans, Rand B. (1992) "Growing Pains." *100 Years: The American Psychological Association: A Historical Perspective*. Washington, DC: American Psychological Association. 76-80.
- Fernberger, Samuel W. (1932). History of the American Psychological Association. *Psychological Bulletin*, 29, 7-15.

share this page:

[FACEBOOK](#)[TWITTER](#)[LINKEDIN](#)[GOOGLE+](#)[EMAIL](#)

Affiliate and International

- [APA Affiliate and International Membership Totals 2001-Present](#)

Performance You Can See & Hear

CONNERS CPT 3
Conners Continuous Performance Test 3*

CONNERS CATA
Conners Continuous Auditory Test of Attention

Visual and Auditory Assessments of Attention

Click for more info

ADVERTISEMENT

[APA Home](#)
[Contact](#)
[Press Room](#)
[Advertise](#)
[APA Store](#)
[Privacy Statement](#)
[Terms of Use](#)
[Access bility](#)
[Website Feedback](#)
[Site Map](#)
[Help](#)

© 2016 American Psychological Association
 750 First St. NE, Washington, DC 20002-4242
 Telephone: (800) 374-2721; (202) 336-5500
 TDD/TTY: (202) 336-6123
[Join APA](#) | [Renew Membership](#)

Follow APA: [f](#) [t](#) [in](#) [G+](#) [v](#) [RSS](#)

More APA websites:

[APA Style](#)
[APA Practice Central](#)
[APA Center for Organizational Excellence](#)
[ACT Raising Safe Kids Program](#)
[APA Education Advocacy Trust](#)
[Online Psychology Laboratory](#)
[Psychology: Science in Action](#)
[APA PsycNET®](#)

JA2847

EXHIBIT 9



FOR: Graduate Students | Divisions | SIGs | OIA

FOLLOW US ON:



- About AERA
- Events & Meetings
- Policy & Advocacy
- Education Research
- Professional Advancement
- Publications
- Membership
- Newsroom

Publications » Books » Standards for Educational & Psychological Testing (2014 Edition)

Publications

Journals

AERA Highlights

Books

Research Points

Online Paper Repository

Online Store

Advertise with AERA

Publications Permissions

Standards for Educational & Psychological Testing (2014 Edition)

ORDER NOW IN THE AERA ONLINE STORE



The 2014 edition of the *Standards for Educational and Psychological Testing* is now on sale. The *Testing Standards* are a product of the American Educational Research Association, the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). Published collaboratively by the three organizations since 1966, it represents the gold standard in guidance on testing in the United States and in many other countries. [Read more](#)

To order the previous edition of *Standards* (1999), please use this [order form](#).

An e-version of the *Testing Standards* is now available. Please see below the pricing options for purchasing an e-book (in e-Pub or e-PDF formats), or a bundle that includes a print volume and a free e-book download.

ISBN 978-0-935302-35-6 (paperback)
ISBN 978-0-935302-41-7 (eBook)

Pricing and Ordering Information:

AERA Members	<p>Print Only: \$49.95 plus shipping Log in to receive your member pricing, then order through the AERA Online Store.</p> <p>E-Book Only: \$49.95 Log in and visit My AERA - Special Member Offers. Copy the member discount code, then follow the link to order the e-Book.</p> <p>Print/e-Book Bundle * \$59.95 Log in to receive your member pricing, then order through the AERA Online Store.</p>
---------------------	---

APA and NCME Print Only:

Capitol Hill Briefing on New Standards



On October 30, AERA filed an amicus curiae brief in the U.S. Supreme Court's reconsideration of *Fisher v. University of Texas at Austin*. [Read press release here](#) and view more information on 2015 efforts [here](#).

Suzanne Lane Discusses Standards

Suzanne Lane Discusses the New Edition of Testing Standards



[Watch on YouTube »](#)

JA2849

Members
\$49.95 plus shipping
Click here to purchase a print copy.

E-Book Only:
\$49.95
Click here to order e-Book only.

Print/e-Book Bundle *
\$59.95
Click here to purchase a print/e-Book bundle.

Non-Member and Institutional Price **

Print Only:
\$69.95 plus shipping
Order now through the [AERA Online Store](#).

E-Book Only:
\$69.95
Click here to order e-Book only.

Print/e-Book Bundle* (The eBook is for single-users only, and is not currently available for institutional purchase)
\$79.95
Order now through the [AERA Online Store](#).

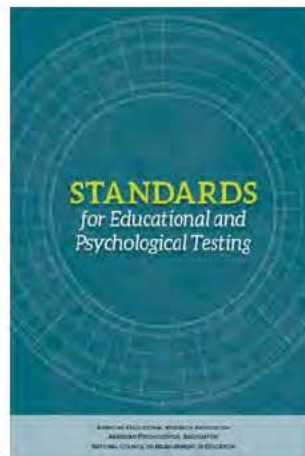
[Mail or Fax Order Form \(PDF\)](#)

***Important Information for Purchasing the Print/e-Book Bundle:**
Print and Bundle orders are fulfilled through the AERA Bookstore. Please [click here](#) for e-Book only orders. After a bundle is purchased, you will receive an email from AERA that includes a link to the e-book sales platform and a coupon for free download. Emails will be sent within 24 hours of receipt during AERA's business hours.

Shipping & Handling:
\$7 for first copy, \$2.00 each additional copy up through 9 copies.

Volume Discount:
** Institutions ordering 10 or more copies will receive a 20% discount off the non-member price. Members may order multiple copies but will not receive an additional discount below the member price. For shipping and handling costs for bulk orders of 10 or more copies, please contact AERA at members@aera.net or 202-238-3200.

AERA Return and Discount Policy
No refunds for returned books. Discounts are not available to agencies.



Testing Standards Hill Briefing Gallery

Search Tags



Testing Standards Hill Briefing Photo Gallery (19 Photos)

JA2850

©2016 American Educational Research Association. All rights reserved.

[Terms Of Use](#) | [Privacy Policy](#) | [Site Map](#) | [Contact Us](#)

1430 K Street NW, Suite 1200, Washington, DC 20005

Phone: (202) 238-3200 | Fax: (202) 238-3250

Designed by [Weber-Shandwick](#) Powered by [eNOAH](#)

JA2851

EXHIBIT 10



[HOME](#) · [BRIEFING ROOM](#) · [SPEECHES & REMARKS](#)

Briefing Room

[Your Weekly Address](#)

Speeches & Remarks

[Press Briefings](#)

[Statements & Releases](#)

[White House Schedule](#)

[Presidential Actions](#)

[Executive Orders](#)

[Presidential Memoranda](#)

[Proclamations](#)

[Legislation](#)

[Pending Legislation](#)

[Signed Legislation](#)

[Vetoed Legislation](#)

[Nominations & Appointments](#)

[Disclosures](#)

The White House

[Office of the Press Secretary](#)

For Immediate Release

January 24, 2012

Remarks by the President in State of the Union Address

JA2853

United States Capitol
Washington, D.C.

9:10 P.M. EST

THE PRESIDENT: Mr. Speaker, Mr. Vice President, members of Congress, distinguished guests, and fellow Americans:

Last month, I went to Andrews Air Force Base and welcomed home some of our last troops to serve in Iraq. Together, we offered a final, proud salute to the colors under which more than a million of our fellow citizens fought -- and several thousand gave their lives.

We gather tonight knowing that this generation of heroes has made the United States safer and more respected around the world. (Applause.) For the first time in nine years, there are no Americans fighting in Iraq. (Applause.) For the first time in two decades, Osama bin Laden is not a threat to this country. (Applause.) Most of al Qaeda's top lieutenants have been defeated. The Taliban's momentum has been broken, and some troops in Afghanistan have begun to come home.

These achievements are a testament to the courage, selflessness and teamwork of America's Armed Forces. At a time when too many of our institutions have let us down, they exceed all expectations. They're not consumed with personal ambition. They don't obsess over their differences. They focus on the mission at hand. They work together.

Imagine what we could accomplish if we followed their example. (Applause.) Think about the

JA2854

USA Case # 735 County that leads the world in educating its people. An America that attracts a new generation of high-tech manufacturing and high-paying jobs. A future where we're in control of our own energy, and our security and prosperity aren't so tied to unstable parts of the world. An economy built to last, where hard work pays off, and responsibility is rewarded.

We can do this. I know we can, because we've done it before. At the end of World War II, when another generation of heroes returned home from combat, they built the strongest economy and middle class the world has ever known. (Applause.) My grandfather, a veteran of Patton's Army, got the chance to go to college on the GI Bill. My grandmother, who worked on a bomber assembly line, was part of a workforce that turned out the best products on Earth.

The two of them shared the optimism of a nation that had triumphed over a depression and fascism. They understood they were part of something larger; that they were contributing to a story of success that every American had a chance to share -- the basic American promise that if you worked hard, you could do well enough to raise a family, own a home, send your kids to college, and put a little away for retirement.

The defining issue of our time is how to keep that promise alive. No challenge is more urgent. No debate is more important. We can either settle for a country where a shrinking number of people do really well while a growing number of Americans barely get by, or we can restore an economy where everyone gets a fair shot, and everyone does their fair share, and everyone plays by the same set of rules. (Applause.) What's at stake aren't Democratic values or Republican values, but American values. And we have to reclaim them.

Let's remember how we got here. Long before the recession, jobs and manufacturing began leaving our shores. Technology made businesses more efficient, but also made some jobs obsolete. Folks at the top saw their incomes rise like never before, but most hardworking Americans struggled with costs that were growing, paychecks that weren't, and personal debt that kept piling up.

In 2008, the house of cards collapsed. We learned that mortgages had been sold to people who couldn't afford or understand them. Banks had made huge bets and bonuses with other people's money. Regulators had looked the other way, or didn't have the authority to stop the bad behavior.

It was wrong. It was irresponsible. And it plunged our economy into a crisis that put millions out of work, saddled us with more debt, and left innocent, hardworking Americans holding the bag. In the six months before I took office, we lost nearly 4 million jobs. And we lost another 4 million before our policies were in full effect.

JA2855

Those are the facts. But for the last 25 months, business has created more than 3 million jobs. (Applause.)

Last year, they created the most jobs since 2005. American manufacturers are hiring again, creating jobs for the first time since the late 1990s. Together, we've agreed to cut the deficit by more than \$2 trillion. And we've put in place new rules to hold Wall Street accountable, so a crisis like this never happens again. (Applause.)

The state of our Union is getting stronger. And we've come too far to turn back now. As long as I'm President, I will work with anyone in this chamber to build on this momentum. But I intend to fight obstruction with action, and I will oppose any effort to return to the very same policies that brought on this economic crisis in the first place. (Applause.)

No, we will not go back to an economy weakened by outsourcing, bad debt, and phony financial profits. Tonight, I want to speak about how we move forward, and lay out a blueprint for an economy that's built to last -- an economy built on American manufacturing, American energy, skills for American workers, and a renewal of American values.

Now, this blueprint begins with American manufacturing.

On the day I took office, our auto industry was on the verge of collapse. Some even said we should let it die. With a million jobs at stake, I refused to let that happen. In exchange for help, we demanded responsibility. We got workers and automakers to settle their differences. We got the industry to retool and restructure. Today, General Motors is back on top as the world's number-one automaker. (Applause.) Chrysler has grown faster in the U.S. than any major car company. Ford is investing billions in U.S. plants and factories. And together, the entire industry added nearly 160,000 jobs.

We bet on American workers. We bet on American ingenuity. And tonight, the American auto industry is back. (Applause.)

What's happening in Detroit can happen in other industries. It can happen in Cleveland and Pittsburgh and Raleigh. We can't bring every job back that's left our shore. But right now, it's getting more expensive to do business in places like China. Meanwhile, America is more productive. A few weeks ago, the CEO of Master Lock told me that it now makes business sense for him to bring jobs back home. (Applause.) Today, for the first time in 15 years, Master Lock's unionized plant in Milwaukee is running at full capacity. (Applause.)

So we have a huge opportunity, at this moment, to bring manufacturing back. But we have to seize it. Tonight, my message to business leaders is simple: Ask yourselves what you can do to bring jobs back to your country, and your country will do everything we can to help you succeed. (Applause.)

JA2856

We should start with our tax code. Right now, companies get tax breaks for moving jobs and profits overseas. Meanwhile, companies that choose to stay in America get hit with one of the highest tax rates in the world. It makes no sense, and everyone knows it. So let's change it.

First, if you're a business that wants to outsource jobs, you shouldn't get a tax deduction for doing it. (Applause.) That money should be used to cover moving expenses for companies like Master Lock that decide to bring jobs home. (Applause.)

Second, no American company should be able to avoid paying its fair share of taxes by moving jobs and profits overseas. (Applause.) From now on, every multinational company should have to pay a basic minimum tax. And every penny should go towards lowering taxes for companies that choose to stay here and hire here in America. (Applause.)

Third, if you're an American manufacturer, you should get a bigger tax cut. If you're a high-tech manufacturer, we should double the tax deduction you get for making your products here. And if you want to relocate in a community that was hit hard when a factory left town, you should get help financing a new plant, equipment, or training for new workers. (Applause.)

So my message is simple. It is time to stop rewarding businesses that ship jobs overseas, and start rewarding companies that create jobs right here in America. Send me these tax reforms, and I will sign them right away. (Applause.)

We're also making it easier for American businesses to sell products all over the world. Two years ago, I set a goal of doubling U.S. exports over five years. With the bipartisan trade agreements we signed into law, we're on track to meet that goal ahead of schedule.

(Applause.) And soon, there will be millions of new customers for American goods in Panama, Colombia, and South Korea. Soon, there will be new cars on the streets of Seoul imported from Detroit, and Toledo, and Chicago. (Applause.)

I will go anywhere in the world to open new markets for American products. And I will not stand by when our competitors don't play by the rules. We've brought trade cases against China at nearly twice the rate as the last administration -- and it's made a difference. (Applause.) Over a thousand Americans are working today because we stopped a surge in Chinese tires. But we need to do more. It's not right when another country lets our movies, music, and software be pirated. It's not fair when foreign manufacturers have a leg up on ours only because they're heavily subsidized.

Tonight, I'm announcing the creation of a Trade Enforcement Unit that will be charged with investigating unfair trading practices in countries like China. (Applause.) There will be more inspections to prevent counterfeit or unsafe goods from crossing our borders. And this Congress should make sure that no foreign company has an advantage over American manufacturing when it comes to accessing financing or new markets like Russia. Our workers are the most productive on Earth, and if the playing field is level, I promise you -- America will

JA2857

I also hear from many business leaders who want to hire in the United States but can't find workers with the right skills. Growing industries in science and technology have twice as many openings as we have workers who can do the job. Think about that -- openings at a time when millions of Americans are looking for work. It's inexcusable. And we know how to fix it.

Jackie Bray is a single mom from North Carolina who was laid off from her job as a mechanic. Then Siemens opened a gas turbine factory in Charlotte, and formed a partnership with Central Piedmont Community College. The company helped the college design courses in laser and robotics training. It paid Jackie's tuition, then hired her to help operate their plant.

I want every American looking for work to have the same opportunity as Jackie did. **Join me in a national commitment to train 2 million Americans with skills that will lead directly to a job.**

(Applause.) My administration has already lined up more companies that want to help. Model partnerships between businesses like Siemens and community colleges in places like Charlotte, and Orlando, and Louisville are up and running. Now you need to give more community colleges the resources they need to become community career centers -- places that teach people skills that businesses are looking for right now, from data management to high-tech manufacturing.

And I want to cut through the maze of confusing training programs, so that from now on, people like Jackie have one program, one website, and one place to go for all the information and help that they need. It is time to turn our unemployment system into a reemployment system that puts people to work. (Applause.)

These reforms will help people get jobs that are open today. But to prepare for the jobs of tomorrow, our commitment to skills and education has to start earlier.

For less than 1 percent of what our nation spends on education each year, we've convinced nearly every state in the country to raise their standards for teaching and learning -- the first time that's happened in a generation.

But challenges remain. And we know how to solve them.

At a time when other countries are doubling down on education, tight budgets have forced states to lay off thousands of teachers. We know a good teacher can increase the lifetime income of a classroom by over \$250,000. A great teacher can offer an escape from poverty to the child who dreams beyond his circumstance. Every person in this chamber can point to a teacher who changed the trajectory of their lives. Most teachers work tirelessly, with modest pay, sometimes digging into their own pocket for school supplies -- just to make a difference.

JA2858

Teachers. So instead of bringing the #1715650 - Filed 01/30/2018 Page 45 of 517 a deal. Give them the resources to keep good teachers on the job, and reward the best ones. (Applause.) And in return, grant schools flexibility: to teach with creativity and passion; to stop teaching to the test; and to replace teachers who just aren't helping kids learn. That's a bargain worth making. (Applause.)

We also know that when students don't walk away from their education, more of them walk the stage to get their diploma. When students are not allowed to drop out, they do better. So tonight, I am proposing that every state -- every state -- requires that all students stay in high school until they graduate or turn 18. (Applause.)

When kids do graduate, the most daunting challenge can be the cost of college. At a time when Americans owe more in tuition debt than credit card debt, this Congress needs to stop the interest rates on student loans from doubling in July. (Applause.)

Extend the tuition tax credit we started that saves millions of middle-class families thousands of dollars, and give more young people the chance to earn their way through college by doubling the number of work-study jobs in the next five years. (Applause.)

Of course, it's not enough for us to increase student aid. We can't just keep subsidizing skyrocketing tuition; we'll run out of money. States also need to do their part, by making higher education a higher priority in their budgets. And colleges and universities have to do their part by working to keep costs down.

Recently, I spoke with a group of college presidents who've done just that. Some schools redesign courses to help students finish more quickly. Some use better technology. The point is, it's possible. So let me put colleges and universities on notice: If you can't stop tuition from going up, the funding you get from taxpayers will go down. (Applause.) Higher education can't be a luxury — it is an economic imperative that every family in America should be able to afford.

Let's also remember that hundreds of thousands of talented, hardworking students in this country face another challenge: the fact that they aren't yet American citizens. Many were brought here as small children, are American through and through, yet they live every day with the threat of deportation. Others came more recently, to study business and science and engineering, but as soon as they get their degree, we send them home to invent new products and create new jobs somewhere else.

That doesn't make sense.

I believe as strongly as ever that we should take on illegal immigration. That's why my administration has put more boots on the border than ever before. That's why there are fewer illegal crossings than when I took office. The opponents of action are out of excuses. We

JA2859

But if election-year politics keeps Congress from acting on a comprehensive plan, let's at least agree to stop expelling responsible young people who want to staff our labs, start new businesses, defend this country. Send me a law that gives them the chance to earn their citizenship. I will sign it right away. (Applause.)

You see, an economy built to last is one where we encourage the talent and ingenuity of every person in this country. That means women should earn equal pay for equal work. (Applause.) It means we should support everyone who's willing to work, and every risk-taker and entrepreneur who aspires to become the next Steve Jobs.

After all, **innovation is what America has always been about**. Most new jobs are created in start-ups and small businesses. So let's pass an agenda that helps them succeed. Tear down regulations that prevent aspiring entrepreneurs from getting the financing to grow. (Applause.) Expand tax relief to small businesses that are raising wages and creating good jobs. Both parties agree on these ideas. So put them in a bill, and get it on my desk this year. (Applause.)

Innovation also demands basic research. Today, the discoveries taking place in our federally financed labs and universities could lead to new treatments that kill cancer cells but leave healthy ones untouched. New lightweight vests for cops and soldiers that can stop any bullet. Don't gut these investments in our budget. Don't let other countries win the race for the future. Support the same kind of research and innovation that led to the computer chip and the Internet; to new American jobs and new American industries.

And nowhere is the promise of innovation greater than in American-made energy. Over the last three years, we've opened millions of new acres for oil and gas exploration, and **tonight, I'm directing my administration to open more than 75 percent of our potential offshore oil and gas resources**. (Applause.) Right now -- right now -- American oil production is the highest that it's been in eight years. That's right -- eight years. **Not only that -- last year, we relied less on foreign oil than in any of the past 16 years**. (Applause.)

But with only 2 percent of the world's oil reserves, oil isn't enough. This country needs an all-out, all-of-the-above strategy that develops every available source of American energy. (Applause.) A strategy that's cleaner, cheaper, and full of new jobs.

We have a supply of natural gas that can last America nearly 100 years. (Applause.) **And my administration will take every possible action to safely develop this energy**. Experts believe this will support more than 600,000 jobs by the end of the decade. And I'm requiring all companies that drill for gas on public lands to disclose the chemicals they use. (Applause.) Because America will develop this resource without putting the health and safety of our citizens at risk.

JA2860

The development of natural gas will create jobs and power trucks and factories that are cleaner and cheaper, proving that we don't have to choose between our environment and our economy. (Applause.) And by the way, it was public research dollars, over the course of 30 years, that helped develop the technologies to extract all this natural gas out of shale rock — reminding us that government support is critical in helping businesses get new energy ideas off the ground. (Applause.)

Now, what's true for natural gas is just as true for clean energy. In three years, our partnership with the private sector has already positioned America to be the world's leading manufacturer of high-tech batteries. Because of federal investments, renewable energy use has nearly doubled, and thousands of Americans have jobs because of it.

When Bryan Ritterby was laid off from his job making furniture, he said he worried that at 55, no one would give him a second chance. But he found work at Energetx, a wind turbine manufacturer in Michigan. Before the recession, the factory only made luxury yachts. Today, it's hiring workers like Bryan, who said, "I'm proud to be working in the industry of the future."

Our experience with shale gas, our experience with natural gas, shows us that the payoffs on these public investments don't always come right away. Some technologies don't pan out; some companies fail. But I will not walk away from the promise of clean energy. I will not walk away from workers like Bryan. (Applause.) I will not cede the wind or solar or battery industry to China or Germany because we refuse to make the same commitment here.

We've subsidized oil companies for a century. That's long enough. (Applause.) It's time to end the taxpayer giveaways to an industry that rarely has been more profitable, and double-down on a clean energy industry that never has been more promising. Pass clean energy tax credits. Create these jobs. (Applause.)

We can also spur energy innovation with new incentives. **The differences in this chamber may be too deep right now to pass a comprehensive plan to fight climate change.** But there's no reason why Congress shouldn't at least set a clean energy standard that creates a market for innovation. So far, you haven't acted. **Well, tonight, I will.** I'm directing my administration to allow the development of clean energy on enough public land to power 3 million homes. And I'm proud to announce that the Department of Defense, working with us, the world's largest consumer of energy, will make one of the largest commitments to clean energy in history — with the Navy purchasing enough capacity to power a quarter of a million homes a year. (Applause.)

Of course, the easiest way to save money is to waste less energy. So here's a proposal: Help manufacturers eliminate energy waste in their factories and give businesses incentives to upgrade their buildings. Their energy bills will be \$100 billion lower over the next decade, and America will have less pollution, more manufacturing, more jobs for construction workers

JA2861

who need them. Send me a bill that creates these jobs. (Applause.)

Building this new energy future should be just one part of a broader agenda to repair America's infrastructure. So much of America needs to be rebuilt. We've got crumbling roads and bridges; a power grid that wastes too much energy; an incomplete high-speed broadband network that prevents a small business owner in rural America from selling her products all over the world.

During the Great Depression, America built the Hoover Dam and the Golden Gate Bridge. After World War II, we connected our states with a system of highways. Democratic and Republican administrations invested in great projects that benefited everybody, from the workers who built them to the businesses that still use them today.

In the next few weeks, I will sign an executive order clearing away the red tape that slows down too many construction projects. But you need to fund these projects. Take the money we're no longer spending at war, use half of it to pay down our debt, and use the rest to do some nation-building right here at home. (Applause.)

There's never been a better time to build, especially since the construction industry was one of the hardest hit when the housing bubble burst. Of course, construction workers weren't the only ones who were hurt. So were millions of innocent Americans who've seen their home values decline. And while government can't fix the problem on its own, responsible homeowners shouldn't have to sit and wait for the housing market to hit bottom to get some relief.

And that's why I'm sending this Congress a plan that gives every responsible homeowner the chance to save about \$3,000 a year on their mortgage, by refinancing at historically low rates. (Applause.) No more red tape. No more runaround from the banks. A small fee on the largest financial institutions will ensure that it won't add to the deficit and will give those banks that were rescued by taxpayers a chance to repay a deficit of trust. (Applause.)

Let's never forget: Millions of Americans who work hard and play by the rules every day deserve a government and a financial system that do the same. It's time to apply the same rules from top to bottom. No bailouts, no handouts, and no copouts. An America built to last insists on responsibility from everybody.

We've all paid the price for lenders who sold mortgages to people who couldn't afford them, and buyers who knew they couldn't afford them. That's why we need smart regulations to prevent irresponsible behavior. (Applause.) Rules to prevent financial fraud or toxic dumping or faulty medical devices -- these don't destroy the free market. They make the free market work better.

There's no question that some regulations are outdated, unnecessary, or too costly. In fact,

JA2862

I've approved fewer regulations in the first three years of my presidency than my Republican predecessor did in his. (Applause.) I've ordered every federal agency to eliminate rules that don't make sense. We've already announced over 500 reforms, and just a fraction of them will save business and citizens more than \$10 billion over the next five years. We got rid of one rule from 40 years ago that could have forced some dairy farmers to spend \$10,000 a year proving that they could contain a spill -- because milk was somehow classified as an oil. With a rule like that, I guess it was worth crying over spilled milk. (Laughter and applause.)

Now, I'm confident a farmer can contain a milk spill without a federal agency looking over his shoulder. (Applause.) Absolutely. But I will not back down from making sure an oil company can contain the kind of oil spill we saw in the Gulf two years ago. (Applause.) I will not back down from protecting our kids from mercury poisoning, or making sure that our food is safe and our water is clean. I will not go back to the days when health insurance companies had unchecked power to cancel your policy, deny your coverage, or charge women differently than men. (Applause.)

And I will not go back to the days when Wall Street was allowed to play by its own set of rules. The new rules we passed restore what should be any financial system's core purpose: Getting funding to entrepreneurs with the best ideas, and getting loans to responsible families who want to buy a home, or start a business, or send their kids to college.

So if you are a big bank or financial institution, you're no longer allowed to make risky bets with your customers' deposits. You're required to write out a "living will" that details exactly how you'll pay the bills if you fail -- because the rest of us are not bailing you out ever again. (Applause.) And if you're a mortgage lender or a payday lender or a credit card company, the days of signing people up for products they can't afford with confusing forms and deceptive practices -- those days are over. Today, American consumers finally have a watchdog in Richard Cordray with one job: To look out for them. (Applause.)

We'll also establish a Financial Crimes Unit of highly trained investigators to crack down on large-scale fraud and protect people's investments. Some financial firms violate major anti-fraud laws because there's no real penalty for being a repeat offender. That's bad for consumers, and it's bad for the vast majority of bankers and financial service professionals who do the right thing. So pass legislation that makes the penalties for fraud count.

And tonight, I'm asking my Attorney General to create a special unit of federal prosecutors and leading state attorney general to expand our investigations into the abusive lending and packaging of risky mortgages that led to the housing crisis. (Applause.) This new unit will hold accountable those who broke the law, speed assistance to homeowners, and help turn the page on an era of recklessness that hurt so many Americans.

Now, a return to the American values of fair play and shared responsibility will help protect our

JA2863

people and our country. But it should also guide us as we look to pay down our debt and invest in our future.

Right now, our most immediate priority is stopping a tax hike on 160 million working Americans while the recovery is still fragile. (Applause.) **People cannot afford losing \$40 out of each paycheck this year.** There are plenty of ways to get this done. So let's agree right here, right now: No side issues. No drama. **Pass the payroll tax cut without delay. Let's get it done.** (Applause.)

When it comes to the deficit, we've already agreed to more than \$2 trillion in cuts and savings. But we need to do more, and that means making choices. Right now, we're poised to spend nearly \$1 trillion more on what was supposed to be a temporary tax break for the wealthiest 2 percent of Americans. Right now, because of loopholes and shelters in the tax code, a quarter of all millionaires pay lower tax rates than millions of middle-class households. Right now, Warren Buffett pays a lower tax rate than his secretary.

Do we want to keep these tax cuts for the wealthiest Americans? Or do we want to keep our investments in everything else — like education and medical research; a strong military and care for our veterans? Because if we're serious about paying down our debt, we can't do both.

The American people know what the right choice is. So do I. As I told the Speaker this summer, I'm prepared to make more reforms that rein in the long-term costs of Medicare and Medicaid, and strengthen Social Security, so long as those programs remain a guarantee of security for seniors.

But in return, we need to change our tax code so that people like me, and an awful lot of members of Congress, pay our fair share of taxes. (Applause.)

Tax reform should follow the Buffett Rule. If you make more than \$1 million a year, you should not pay less than 30 percent in taxes. And my Republican friend Tom Coburn is right: Washington should stop subsidizing millionaires. In fact, if you're earning a million dollars a year, you shouldn't get special tax subsidies or deductions. On the other hand, if you make under \$250,000 a year, like 98 percent of American families, your taxes shouldn't go up. (Applause.) You're the ones struggling with rising costs and stagnant wages. You're the ones who need relief.

Now, you can call this class warfare all you want. But asking a billionaire to pay at least as much as his secretary in taxes? Most Americans would call that common sense.

We don't begrudge financial success in this country. We admire it. When Americans talk about folks like me paying my fair share of taxes, it's not because they envy the rich. It's because they understand that when I get a tax break I don't need and the country can't afford,

JA2864

a fixed income, or a student trying to get through school, or a family trying to make ends meet. That's not right. Americans know that's not right. They know that this generation's success is only possible because past generations felt a responsibility to each other, and to the future of their country, and they know our way of life will only endure if we feel that same sense of shared responsibility. That's how we'll reduce our deficit. That's an America built to last. (Applause.)

Now, I recognize that people watching tonight have differing views about taxes and debt, energy and health care. But no matter what party they belong to, I bet most Americans are thinking the same thing right about now: Nothing will get done in Washington this year, or next year, or maybe even the year after that, because Washington is broken.

Can you blame them for feeling a little cynical?

The greatest blow to our confidence in our economy last year didn't come from events beyond our control. It came from a debate in Washington over whether the United States would pay its bills or not. Who benefited from that fiasco?

I've talked tonight about the deficit of trust between Main Street and Wall Street. But the divide between this city and the rest of the country is at least as bad -- and it seems to get worse every year.

Some of this has to do with the corrosive influence of money in politics. So together, let's take some steps to fix that. **Send me a bill that bans insider trading by members of Congress; I will sign it tomorrow.** (Applause.) Let's limit any elected official from owning stocks in industries they impact. Let's make sure people who bundle campaign contributions for Congress can't lobby Congress, and vice versa -- an idea that has bipartisan support, at least outside of Washington.

Some of what's broken has to do with the way Congress does its business these days. A simple majority is no longer enough to get anything -- even routine business -- passed through the Senate. (Applause.) Neither party has been blameless in these tactics. Now both parties should put an end to it. (Applause.) **For starters, I ask the Senate to pass a simple rule that all judicial and public service nominations receive a simple up or down vote within 90 days.** (Applause.)

The executive branch also needs to change. Too often, it's inefficient, outdated and remote. (Applause.) That's why I've asked this Congress to grant me the authority to consolidate the federal bureaucracy, so that our government is leaner, quicker, and more responsive to the needs of the American people. (Applause.)

Finally, none of this can happen unless we also lower the temperature in this town. We need

JA2865

to end the war; that the war is not a perpetual campaign of mutual destruction; that politics is about clinging to rigid ideologies instead of building consensus around common-sense ideas.

I'm a Democrat. **But I believe what Republican Abraham Lincoln believed: That government should do for people only what they cannot do better by themselves, and no more.**

(Applause.) That's why my education reform offers more competition, and more control for schools and states. That's why we're getting rid of regulations that don't work. That's why our health care law relies on a reformed private market, not a government program.

On the other hand, even my Republican friends who complain the most about government spending have supported federally financed roads, and clean energy projects, and federal offices for the folks back home.

The point is, we should all want a smarter, more effective government. And while we may not be able to bridge our biggest philosophical differences this year, we can make real progress. With or without this Congress, I will keep taking actions that help the economy grow. But I can do a whole lot more with your help. Because when we act together, there's nothing the United States of America can't achieve. (Applause.) That's the lesson we've learned from our actions abroad over the last few years.

Ending the Iraq war has allowed us to strike decisive blows against our enemies. From Pakistan to Yemen, the al Qaeda operatives who remain are scrambling, knowing that they can't escape the reach of the United States of America. (Applause.)

From this position of strength, we've begun to wind down the war in Afghanistan. Ten thousand of our troops have come home. Twenty-three thousand more will leave by the end of this summer. This transition to Afghan lead will continue, and we will build an enduring partnership with Afghanistan, so that it is never again a source of attacks against America. (Applause.)

As the tide of war recedes, a wave of change has washed across the Middle East and North Africa, from Tunis to Cairo; from Sana'a to Tripoli. A year ago, Qaddafi was one of the world's longest-serving dictators -- a murderer with American blood on his hands. Today, he is gone. **And in Syria, I have no doubt that the Assad regime will soon discover that the forces of change cannot be reversed, and that human dignity cannot be denied.** (Applause.)

How this incredible transformation will end remains uncertain. But we have a huge stake in the outcome. And while it's ultimately up to the people of the region to decide their fate, we will advocate for those values that have served our own country so well. We will stand against violence and intimidation. We will stand for the rights and dignity of all human beings -- men and women; Christians, Muslims and Jews. We will support policies that lead to strong and stable democracies and open markets, because tyranny is no match for liberty.

JA2866

And we will safeguard America's own security against those who threaten our citizens, our friends, and our interests. Look at Iran. Through the power of our diplomacy, a world that was once divided about how to deal with Iran's nuclear program now stands as one. The regime is more isolated than ever before; its leaders are faced with crippling sanctions, and as long as they shirk their responsibilities, this pressure will not relent.

Let there be no doubt: America is determined to prevent Iran from getting a nuclear weapon, and I will take no options off the table to achieve that goal. (Applause.)

But a peaceful resolution of this issue is still possible, and far better, and if Iran changes course and meets its obligations, it can rejoin the community of nations.

The renewal of American leadership can be felt across the globe. Our oldest alliances in Europe and Asia are stronger than ever. Our ties to the Americas are deeper. Our ironclad commitment -- and I mean ironclad -- to Israel's security has meant the closest military cooperation between our two countries in history. (Applause.)

We've made it clear that America is a Pacific power, and a new beginning in Burma has lit a new hope. From the coalitions we've built to secure nuclear materials, to the missions we've led against hunger and disease; from the blows we've dealt to our enemies, to the enduring power of our moral example, America is back.

Anyone who tells you otherwise, anyone who tells you that America is in decline or that our influence has waned, doesn't know what they're talking about. (Applause.)

That's not the message we get from leaders around the world who are eager to work with us. That's not how people feel from Tokyo to Berlin, from Cape Town to Rio, where opinions of America are higher than they've been in years. Yes, the world is changing. No, we can't control every event. But America remains the one indispensable nation in world affairs -- and as long as I'm President, I intend to keep it that way. (Applause.)

That's why, working with our military leaders, I've proposed a new defense strategy that ensures we maintain the finest military in the world, while saving nearly half a trillion dollars in our budget. To stay one step ahead of our adversaries, I've already sent this Congress legislation that will secure our country from the growing dangers of cyber-threats. (Applause.)

Above all, our freedom endures because of the men and women in uniform who defend it. (Applause.) As they come home, we must serve them as well as they've served us. That includes giving them the care and the benefits they have earned -- which is why we've increased annual VA spending every year I've been President. (Applause.) And it means enlisting our veterans in the work of rebuilding our nation.

With the bipartisan support of this Congress, we're providing new tax credits to companies that

JA2867

hire vets. Michelle and Jill Biden have worked with American businesses to secure a pledge of 135,000 jobs for veterans and their families. And tonight, I'm proposing a Veterans Jobs Corps that will help our communities hire veterans as cops and firefighters, so that America is as strong as those who defend her. (Applause.)

Which brings me back to where I began. Those of us who've been sent here to serve can learn a thing or two from the service of our troops. When you put on that uniform, it doesn't matter if you're black or white; Asian, Latino, Native American; conservative, liberal; rich, poor; gay, straight. When you're marching into battle, you look out for the person next to you, or the mission fails. When you're in the thick of the fight, you rise or fall as one unit, serving one nation, leaving no one behind.

One of my proudest possessions is the flag that the SEAL Team took with them on the mission to get bin Laden. On it are each of their names. Some may be Democrats. Some may be Republicans. But that doesn't matter. Just like it didn't matter that day in the Situation Room, when I sat next to Bob Gates -- a man who was George Bush's defense secretary -- and Hillary Clinton -- a woman who ran against me for president.

All that mattered that day was the mission. No one thought about politics. No one thought about themselves. One of the young men involved in the raid later told me that he didn't deserve credit for the mission. It only succeeded, he said, because every single member of that unit did their job -- the pilot who landed the helicopter that spun out of control; the translator who kept others from entering the compound; the troops who separated the women and children from the fight; the SEALs who charged up the stairs. More than that, the mission only succeeded because every member of that unit trusted each other -- because you can't charge up those stairs, into darkness and danger, unless you know that there's somebody behind you, watching your back.

So it is with America. Each time I look at that flag, I'm reminded that our destiny is stitched together like those 50 stars and those 13 stripes. No one built this country on their own. This nation is great because we built it together. This nation is great because we worked as a team. This nation is great because we get each other's backs. And if we hold fast to that truth, in this moment of trial, there is no challenge too great; no mission too hard. As long as we are joined in common purpose, as long as we maintain our common resolve, our journey moves forward, and our future is hopeful, and the state of our Union will always be strong.

Thank you, God bless you, and God bless the United States of America. (Applause.)

END

10:16 P.M. EST

JA2868

Learn more

- Take a **deep dive** into the data behind the President's plan
- Find out how you can **talk to Obama Administration officials** about the President's plan
- Watch the **enhanced version** of the 2012 State of the Union Address
- Video: Go **behind the scenes** as the President prepared his speech
- Photo Gallery: **Scenes from the State of the Union**
- Interactive: **Who joined the First Lady for the speech?**

[HOME](#)

[BRIEFING ROOM](#)

[ISSUES](#)

[THE ADMINISTRATION](#)

[PARTICIPATE](#)

[1600 PENN](#)

[En Español](#)

[Accessibility](#)

[Copyright Information](#)

[Privacy Policy](#)

[USA.gov](#)

JA2869

EXHIBIT 11



HOME · BLOG

Everything You Need to Know: Waivers, Flexibility, and Reforming No Child Left Behind

FEBRUARY 9, 2012 AT 6:21 PM ET BY MEGAN SLACK



Summary: We've put together a quick primer to help you understand the details behind President Obama's announcement that 10 states will receive waivers exempting them from meeting No Child Left Behind's most troublesome and restrictive requirements in exchange for setting their own higher, more honest standards for student success.



President Barack Obama, with Secretary of Education Arne Duncan, delivers remarks on education reform and the

JA2871

Explaining that our kids can't wait any long for Congress to act, [President Barack Obama announced today](#) that ten states that have agreed to implement bold education reforms will receive waivers from the burdensome mandates of the federal education law known as No Child Left Behind. These [waivers will give states the flexibility needed](#) to raise student achievement standards, improve school accountability, and increase teacher effectiveness. The ten states approved for flexibility are Colorado, Florida, Georgia, Indiana, Kentucky, Massachusetts, Minnesota, New Jersey, Oklahoma, and Tennessee. (UPDATE: An eleventh state, New Mexico, was also [approved for a waiver](#) shortly after the initial announcement).

So what does all this mean for our schools? What's the problem with No Child Left Behind? What's a waiver anyway, and why do states need flexibility? To answer these questions, we've put together a quick primer to help you understand the details behind today's announcement.

WHAT'S THE DEAL WITH NO CHILD LEFT BEHIND?

No Child Left Behind, the most current version of the Elementary and Secondary Education Act, was signed into law in 2001—and is five years overdue to be re-written by Congress. The law's objective was admirable. It shined light on achievement gaps and increased accountability at the school level for high-need students. And there's no question that setting goals and holding schools accountable for meeting them is central to an education system that prepares students to compete in a global, 21st century economy.

As written, however, No Child Left Behind has serious flaws. In fact, some of the law's requirements are actually stifling the kind of reforms we need to really improve student achievement, teacher effectiveness, and school accountability. For example, it determines whether schools are falling behind based on test scores. It imposes punitive labels and prescribes one-size-fits-all federal mandates for fixing failing schools. It's led states to narrow curriculum to focus more on teaching to the test and less on teaching everything else student need to know, and to lower standards to make them easier to meet

The Obama administration has worked extensively with Congress to re-write the law, and even submitted its own blueprint for education reform in March 2010, but legislators have not moved forward.

WHAT ARE WAIVERS AND WHAT DO THEY HAVE TO DO WITH NO CHILD LEFT BEHIND?

Waivers provide an opportunity to fix what's wrong with No Child Left Behind without waiting any longer for Congress to Act. States receiving waivers are given flexibility that exempts them from meeting the law's most troublesome and restrictive requirements in exchange for setting their own higher, more honest standards for student success.

For example, waivers will give states the flexibility to:

- Set their own ambitious but achievable terms for closing achievement gaps and ensuring students

JA2872

are not at least 90 percent proficient in reading and math, instead of 85 percent. The goal is to cut the number of students that are not at least 90 percent proficient by 2014. Kentucky, for example, has set a goal to cut the number of underperforming students in half over the next five years.

- Design their own strategies to improve their lowest-performing schools and measure student progress year over year, instead of relying on absolute numbers and a federally prescribed, “one size fits all” approach. Colorado, for example, another state receiving a waiver, is launching a website that will allow teachers and parents can see exactly how much progress students are making, and how different schools measure up.

WHY DO STATES NEED FLEXIBILITY?

States need the flexibility to move forward with innovative education reforms they design themselves — rather than a federal mandate—without sacrificing high standards or lowering accountability. After all, what works for Kentucky doesn’t necessarily work for New Jersey, and the parents and educators who live and work in each place are best-positioned to know the needs of their own communities.

There is still no clear bipartisan path in Congress for ESEA reauthorization – and we can’t wait any longer. Schools and districts continue their daily work of educating students, while also planning for next school year, and states need this flexibility now to implement plans for reform and improvement. Today’s announcement continues a process the President announced last September.

The fact is, most states are already pursuing reforms that go above and beyond the requirements in No Child Left Behind, and waivers will help them continue that progress. More than 40 states have adopted common standards that define what it means to be college and career ready, just as many have designed assessments to measure student progress toward achieving those standards. States have reformed teacher and principal evaluations to better determine which ones are effective and which ones aren’t, and developed support systems to help the less effective ones improve.

HOW DID THESE STATES QUALIFY FOR WAIVERS?

President Obama offered every state a deal: If you’re willing to set higher, more honest standards based on a clear goal that every student can graduate ready for college or a career, we’ll give you the flexibility to meet those standards.

In addition to setting new performance targets for student achievement, states had to prove that they were serious by developing a plan addressing three critical areas:

- **Preparing students for college and careers:** States must have already adopted college- and career-ready standards in reading and math that raise the achievement of all students, including English language Learners and students with disabilities. Additionally, states must create a plan to help schools and districts implement those standards and administer statewide tests to measure progress.
- **Hold schools accountable for making progress:** States must establish an accountability system

JA2873

that recognizes and rewards both high-performing schools as well as those that are making significant gains in improving student achievement. And they must develop targeted strategies to turn around the lowest performing schools and help groups of students with the greatest needs.

- **Improving teacher and principal effectiveness:** States must set guidelines for teacher and principal evaluation and support systems, developed with input from educators and principals. Evaluation systems should assess performance using factors beyond test scores—such as principal observation, peer review, student work, or parent and student feedback—and provide teachers with both constructive advice for improving and support in doing so.

WHAT'S NEXT?

Just as the administration worked extensively with Congress to try re-write No Child Left Behind before announcing last September that it would offer states flexibility waivers, President Obama will continue to call on Congress to reform the law while offering states that are willing to set higher standards for their students the chance to do so.

In fact, in addition to the 10 states that requested the flexibility to implement reforms through this initial round of waivers, an 11th application is still being revised and reviewed, and 28 other states along with Puerto Rico and the District of Columbia have also expressed interest in receiving waivers.

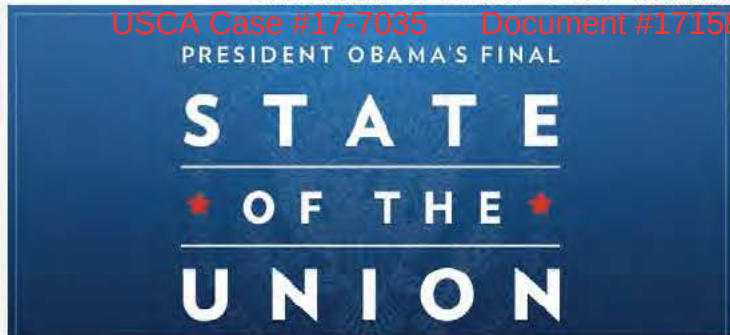
As President Obama explained this afternoon, “if we’re serious about helping our children reach their potential, the best ideas aren’t going to come from Washington alone. Our job is to harness those ideas, and to hold states and schools accountable for making them work.”

Update: On May 29, 2012 the U.S. Department of Education granted waivers to an additional eight states: Connecticut, Delaware, Louisiana, Maryland, New York, North Carolina, Ohio, and Rhode Island, which brings the total number of states to receive waivers to 19, with an additional 18 applications still under review.



Megan Slack

Former Deputy Director of Digital Content for the Office of Digital Strategy



THE FINAL STATE OF THE UNION

Watch President Obama's final State of the Union address.

EXPLORE PAST ADDRESSES

Check out past State of the Union speeches now annotated on Genius.



CLEAN POWER PLAN

Learn more about the biggest step we've ever taken to combat climate change.



HOME BRIEFING ROOM ISSUES THE ADMINISTRATION PARTICIPATE 1600 PENN

[En Español](#) | [Accessibility](#) | [Copyright Information](#) | [Privacy Policy](#) | [USA.gov](#)

EXHIBIT 12



CRIMINAL JURY INSTRUCTIONS FOR THE...

\$186.00 LexisNexis



Local

National resolution against high-stakes tests released



32

By Valerie Strauss April 24, 2012 [Follow @valenestrauss](#)

A [national resolution protesting high-stakes standardized testing](#) was released Tuesday by a coalition of national education, civil rights and parents groups, as well as educators who are trying to build a broad-based movement against the Obama administration's test-centric school reform program.

This is the [latest in a series of recent initiatives taken](#) around the country by academics, educators, parents and others to protest the use of student standardized test scores for high-stakes decisions, including teacher and principal evaluation, student grade promotion and high school graduation.

The high-stakes testing era started with the advent of No Child Left Behind in 2002, and though NCLB has largely been discredited, the Obama administration's policies have expanded the use of test scores as assessment tools not only for students, but also for teachers and principals.

Many researchers in the assessment field have warned against using standardized test scores for high-stakes decisions, saying they are unreliable for such a purpose. High-stakes standardized testing, they say, has led to the narrowing of the curriculum; classrooms where "teaching to the test" is paramount; and unfair evaluation of students, teachers, principals and schools.

The resolution (see text below) is modeled on one passed in recent months by



Most Read

- 1 [Blizzard watch: Severe snowstorm likely Friday through Sunday](#)
- 2 [How much snow are local forecasters and computer models predicting?](#)
- 3 [How to prepare for this weekend's high-impact winter storm](#)
- 4 [D.C. area forecast: Serious late week winter storm; Winter weather advisory for snow during PM rush today](#)

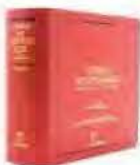
JA2877

more than 300 school boards in Texas, where the Republican state education commissioner, Robert Scott, made news in February by saying the mentality that standardized testing is the “end-all, be-all” is a “perversion” of what a quality education should be, and calling “the assessment and accountability regime” not only “a cottage industry but a military-industrial complex.”

The organizers want organizations and individuals [to endorse the resolution](#), which asks officials in every state to “reexamine public school accountability systems” and to “develop a system based on multiple forms of assessment which does not require extensive standardized testing” and “more accurately reflects the broad range of student learning.”

The resolution also calls on Congress and the Obama administration to rewrite the Elementary and Secondary Education Act, the federal education law known in its current form as No Child Left Behind, in a way that reduces the mandate for standardized tests, promotes multiple forms of evidence that students are learning and does not mandate that student test scores be used to evaluate educators.

Criminal Jury Instructions for the...



\$186.00
Shop this and many other great items today at the LexisNexis® Store.
LexisNexis



“Parents are fed up with constant testing,” Pamela Grundy, of Parents Across America, was quoted as saying in a statement. She helped lead a community revolt against expanding testing in Charlotte, N.C., last year. “We want our elected leaders to support real learning, not endless evaluation,” she said.

The national resolution was written by Advancement Project; Asian American Legal Defense and Education Fund; FairTest; Forum for Education and Democracy; MecklenburgACTS; Deborah Meier; NAACP Legal Defense and Educational Fund, Inc.; National Education Association;

JA2878

The suggestion box is open — it's time to name this winter storm!

Unlimited Access to The Post. Just 99¢

The Most Popular All Over

The Baltimore Sun
One officers statements ruled admissible in Freddie Gray case, as 3 others withdraw efforts to block

International Business Times
Astronomers: Evidence solar system has dark mysterious Ninth Planet 10,000 times the mass of Pluto

WBUR (Boston)
Forecast: Possibly Historic Blizzard To Hit Boston Monday Into Tuesday

The Most Popular stories around the web

New York Performance Standards Consortium, Tracy Novick, Parents Across America; Parents United for Responsible Education - Chicago; Diane Ravitch; Race to Nowhere; Time Out From Testing; and United Church of Christ Justice and Witness Ministries.

Already a number of other organizations and individuals from around the country [have signed on to the resolution](#).

In recent months, protests by parents and educators have been increasing in a number of states in addition to Texas, including New York, California and Illinois. This resolution is an effort to make a national statement about the dangers of high-stakes testing that gets the attention of policy makers at the state and federal levels.

Here's the text of the national resolution:

WHEREAS, our nation's future well-being relies on a high-quality public education system that prepares all students for college, careers, citizenship and lifelong learning, and strengthens the nation's social and economic well-being; and

WHEREAS, our nation's school systems have been spending growing amounts of time, money and energy on high-stakes standardized testing, in which student performance on standardized tests is used to make major decisions affecting individual students, educators and schools; and

WHEREAS, the over-reliance on high-stakes standardized testing in state and federal accountability systems is undermining educational quality and equity in U.S. public schools by hampering educators' efforts to focus on the broad range of learning experiences that promote the innovation, creativity, problem solving, collaboration, communication, critical thinking and deep subject-matter knowledge that will allow students to thrive in a democracy and an increasingly global society and economy; and

WHEREAS, it is widely recognized that standardized testing is an inadequate and often unreliable measure of both student learning and educator effectiveness; and

WHEREAS, the over-emphasis on standardized testing has caused

Our Online Games

Play right from this page



Mahjongg Dimensions

Genre(s): [Strategy](#)

It's 3D Mahjongg- you don't even need to wear 3D glasses!



The Sunday Crossword by Evan Birnholz

Genre(s): [Word](#)

Online crossword.



Spider Solitaire

Genre(s): [Card](#)

Spider Solitaire is known as the king of all solitaire games!



Daily Crossword

Genre(s): [Word](#)

Challenge your crossword skills everyday with a huge variety of puzzles waiting for you to solve.

Get the Local Headlines newsletter

Daily headlines about the Washington region.

washingtonpost.com

© 1996-2016 The Washington Post

[Help and Contact Us](#)

[Terms of Service](#)

[Privacy Policy](#)

[Print Products Terms of Sale](#)

[Digital Products Terms of Sale](#)

[Submissions and Discussion Policy](#)

[RSS Terms of Service](#)

[Ad Choices](#)

JA2879

considerable damage to our schools, including narrowing the curriculum, teaching to the test, reducing love of learning, pushing students out of school, driving excellent teachers out of the profession, and undermining school climate; and

WHEREAS, high-stakes standardized testing has negative effects for students from all backgrounds, and especially for low-income students, English language learners, children of color, and those with disabilities; and

WHEREAS, the culture and structure of the systems in which students learn must change in order to foster engaging school experiences that promote joy in learning, depth of thought and breadth of knowledge for students; therefore be it

RESOLVED, that [your organization name] calls on the governor, state legislature and state education boards and administrators to reexamine public school accountability systems in this state, and to develop a system based on multiple forms of assessment which does not require extensive standardized testing, more accurately reflects the broad range of student learning, and is used to support students and improve schools; and

RESOLVED, that [your organization name] calls on the U.S. Congress and Administration to overhaul the Elementary and Secondary Education Act, currently known as the “No Child Left Behind Act,” reduce the testing mandates, promote multiple forms of evidence of student learning and school quality in accountability, and not mandate any fixed role for the use of student test scores in evaluating educators.

-0-



JA2880

www.washingtonpost.com/blogs/answer-sheet.



Valerie Strauss covers education and runs The Answer Sheet blog.

Share on Facebook

Share on Twitter

32 Comments

PAID PROMOTED STORIES

Recommended by
Outbrain



What I Learned From The Guy Who Trains Navy SEALs

The Blog of Author Tim Ferriss



Do You Have Royal Blood? Your Last Name May Tell You.

Ancestry



How Older Men Tighten Their Skin

The Modern Man Today



These 4 credit cards offer out-of-this-world cash back deals

NextAdvisor



Made in the USA isn't a gimmick. It's the Gillette promise and what those

Gillette Shave Club



This Is The Before Photo, You Won't Believe The After

HGTV

32 Show Comments

ozobot



PRE-ORDER NOW!

PROMOCODE: OZOBOT10



JA2881