CLEAR Exam Review is a journal, published twice a year, reviewing issues affecting testing and credentialing. CER is published by the Council on Licensure, Enforcement, and Regulation, 403 Marquis Ave., Suite 200, Lexington, KY 40502.

Design and composition of this journal have been underwritten by Prometric, which specializes in the design, development, and full-service operation of high-quality licensing, certification and other adult examination programs.

Subscriptions to CER are sent free of charge to all CLEAR members and are available for $30 per year to others. Contact CLEAR at (859) 269-1289 or cer@clearhq.org for membership and subscription information.

Advertisements and Classified (e.g., position vacancies) for CER may be reserved by contacting CLEAR at the address or phone number noted above. Ads are limited in size to 1/4 or 1/2 page, and cost $100 or $200, respectively, per issue.

Editorial Board
Steven Nettles
Applied Measurement Professionals
Jim Zukowski
360training

Coeditor
Elizabeth Witt, Ph.D.
Witt Measurement Consulting
Laingsburg, MI
WittMeasure@aol.com

Coeditor
Sandra Greenberg, Ph.D.
Professional Examination Service
New York, NY
sgreenberg@proexams.org

Copyright ©2015 Council on Licensure, Enforcement, and Regulation. All rights reserved. ISSN 1076-8025

Contents

FROM THE EDITORS ................................................................. 1
Sandra Greenberg, Ph.D.
Elizabeth Witt, Ph.D.

ARTICLES
A User’s Guide to the 2014 Joint Testing Standards for Credentialing Programs
Ronald C. Rodgers, Ph.D.

A Review of the Newly Adopted NCCA Standards
Brian D. Bontempo, Ph.D.

Subscore reporting: An updated reminder from the revised Standards for Educational and Psychological Testing
Corina M. Owens, Ph.D. and Vincent Lima

COLUMNS
Abstracts and Updates ....................................................... 21
George T. Gray, Ed.D.

Legal Beat ................................................................. 28
Dale J. Atkinson, Esq.

Perspectives on Testing ..................................................... 30
Chuck Friedman, Ph.D.
From the Editors

Welcome to the Spring 2015 issue of the Clear Exam Review, volume XXV, number 1. Spring is a time of renewal, and this issue features several articles related to the long-awaited renewal and revision of various testing standards. The Standards for Educational and Psychological Testing (Standards), which serves as a guide for sound practice and fairness in high-stakes testing, was recently updated to reflect current technologies and other advances in assessment. In addition, the Standards for the Accreditation of Certification Programs (NCCA Standards) published by the National Commission for Certifying Agencies (NCCA) has also been revised, with the new NCCA Standards to take effect in 2016. The NCCA Standards are used primarily in accrediting certification programs, but the principles they express are relevant in licensure as well.

Ronald Rodgers has created a User’s Guide to the 2014 Joint Testing Standards for Credentialing Programs that incorporates input from key members of the Joint Committee of the AERA, APA, and NCME charged with the task of updating the Standards. This article reviews the standards that are most relevant in the context of credentialing and highlights differences between the 2014 and 1999 editions of the Standards. Readers will undoubtedly find this article valuable in navigating the revised Standards.

In A Review of the Newly Adopted NCCA Standards, Brian Bontempo, who served on the Revisions Steering Committee for the NCCA Standards, describes the changes to the NCCA Standards, explains the relevance of this document to licensure programs, describes the revision process, and highlights changes from the earlier edition. As the NCCA Standards reflects current thinking regarding sound practice in credentialing programs, we believe that readers will find this article useful in identifying what the professional community expects of both licensure and certification programs.

Returning to the AERA, APA, and NCME Standards, the article by Owens and Lima describes the chapter on workplace testing and credentialing and then focuses on the application of a revised standard related to subscore reporting. Reliability issues are discussed, and several methods of providing feedback to failing candidates are illustrated.

George Gray’s Abstracts and Updates includes a review of the recent issue of Educational Measurement: Issues and Practice that was devoted to the revised Standards. George also reviews a variety of recent publications, including a textbook on item development and articles covering a range of topics: a comparison of IRT and classical test theory; two illustrations of exam development and validation via job analysis; two articles addressing standard setting; a study of score gains among repeat examinees; two articles addressing the validity of a medical certification exam; a study of the relationship of certification and safety knowledge in the food industry; an investigation of the predictive validity of a medical in-training examination; an evaluation of the effects of interrupting a computer-delivered test; an assessment of the effects of
candidate strategies on automated scoring; and two related articles on the reliability and value of subscores.

In the Legal Beat column, Dale Atkinson describes a case in which one exam preparation organization sued another over the harvesting of copyrighted items. The outcome of this case may have implications for licensure and certification testing programs as well.

Perspectives on Testing, presented by Chuck Friedman, provides expert responses to questions posed by CLEAR members. Questions addressed in this issue focus on continuing competence and feedback to candidates. Readers are encouraged to submit questions for future columns and/or CLEAR conference sessions to clear@clearhq.org. Please enter “CER Perspectives on Testing” in the subject line.

Brian Bontempo’s Technology and Testing column will return in the Fall 2015 issue of the CLEAR Exam Review, continuing the series on data visualization.

This issue includes three feature articles about newly revised testing standards and three regularly scheduled columns. We are grateful to the authors and reviewers for the invaluable contributions they have made to the credentialing community.

Read on, learn, and enjoy…
Publication of the 2014 Standards for Educational and Psychological Testing provides credentialing programs a fresh opportunity to assess whether their test development, administration and candidate procedures fully address the needs of their candidates and sponsoring boards. That may very well be the primary benefit many programs will realize from a careful review of the new Standards.

A 15-member Joint Committee comprised of nominees from AERA, APA and NCME was created in 2004. Their instructions were to update the Standards rather than to recreate them. Wayne Camara, chair of the project’s Management Committee, and Nancy Tippins, a member of the Joint Committee, summarized their assignment in a review of the 2014 Standards at the Association of Test Publishers (ATP) conference in Palm Springs, CA in March 2015:

- Fairness – To update ideas on universal design and accommodations and to expand protected groups to include age-defined groups.
- Technology – To reflect the impact of technology on test formats, content, administration and reporting.
- Accountability – To focus on educational accountability in the use of testing.
- Employment Testing and Credentialing – To reflect changes in practice and delivery of employment-related examinations for hiring, licensure and certification activities.
- Formatting and Prioritization – To consider organization of the Standards to reflect an “overarching concept” for all chapters in Part I (Foundations) and Part II (Operations).

AERA published the product of these efforts mid-year in 2014. Table 1 displays organization of the 1999 Standards and the 2014 Standards. The 2014 version consolidates content in some areas and renames six chapters (highlighted) to clarify their intent, but most of the changes are embedded in the details of the new Standards, as the Joint Committee focused on updating over recreating the Joint Standards.

Table 1 also documents primary organizational changes in the 2014 Standards. Part I is renamed Foundations and consists of the first three chapters titled Validity, Reliability/Precision and Errors of Measurement, and Fairness in Testing. Three chapters on Fairness
from the 1999 Standards were consolidated into a single chapter in 2014. Part II unites six chapters under the heading Operations and picks up the chapter on Rights and Responsibilities of Test Users. Part III retains the focus on Testing Applications.

Significant changes in chapter headings are highlighted in Table 1. Each of these will be addressed in more detail later in this review of the Standards.

**Table 1. Comparison of 1999 Standards to 2014 Standards**

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Validity</td>
<td>1. Validity</td>
</tr>
<tr>
<td>2. Reliability and Errors of Measurement</td>
<td>2. Reliability/Precision and Errors of Measurement</td>
</tr>
<tr>
<td>3. Test Development and Revision</td>
<td>3. Fairness in Testing</td>
</tr>
<tr>
<td>4. Scales, Norms and Score Comparability</td>
<td></td>
</tr>
<tr>
<td>5. Test Administration, Scoring and Reporting</td>
<td></td>
</tr>
<tr>
<td>6. Supporting Documentation for Tests</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part II: Fairness in Testing</th>
<th>Part II: Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>7. Fairness in Testing and Test Use</td>
<td>4. Test Design and Development</td>
</tr>
<tr>
<td>8. The Rights and Responsibilities of Test Takers</td>
<td>5. Scores, Scales, Norms, Score Linking and Cut Scores</td>
</tr>
<tr>
<td>9. Testing Individuals of Diverse Linguistic Backgrounds</td>
<td>6. Test Administration, Scoring, Reporting and Interpretation</td>
</tr>
<tr>
<td>10. Testing Individuals with Disabilities</td>
<td>7. Supporting Documentation for Tests</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part III: Testing Applications</th>
<th>Part III: Testing Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>11. The Responsibilities of Test Users</td>
<td>10. Psychological Testing and Assessment</td>
</tr>
<tr>
<td>15. Testing in Program Evaluation and Public Policy</td>
<td></td>
</tr>
</tbody>
</table>

Source: Camara & Tippins (2015).

**Part I Foundations**

Individual standards are organized and presented differently in the 2014 version to help users identify common themes and clusters of related standards. Standards in Foundations (Part I) and Operations (Part II) are introduced by an “overarching standard designed to convey the intent of the chapter” (2014 Standards, p. 7). For example, Standard 1.0 in Validity specifies:

“Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided.” (2014 Standards, p. 23)

This helps clarify the guiding principle of Chapter 1. It is followed by three thematic clusters of standards on validity:

1. Establishing Intended Uses and Interpretations
2. Issues Regarding Samples and Settings Used in Validation
3. Specific Forms of Validity Evidence

This organizational approach continues throughout Parts I and II of the 2014 Standards. It should help users understand and respond to the general intent and thematic groups of related standards. It also simplifies the task of locating standards most likely to apply to a testing program’s problems, needs and circumstances. It should help make the 2014 Standards a much more user-friendly resource.
KEY ELEMENTS IN CHAPTER 1 REFLECT THE MOST ESSENTIAL IMPACT OF THE 2014 STANDARDS FOR CREDENTIALING PROGRAMS BASED ON TWO PRIMARY PRINCIPLES IN THE VALIDITY CHAPTER:

- “Validity of test score interpretations depends not only on the uses of test scores but specifically on the claims that underlie the theory of action for these uses” (2014 Standards, pp. 19-20).
- “Standards of evidence may vary with the stakes involved in the use or interpretation of the test scores. Higher stakes may entail higher standards of evidence” (2014 Standards, p. 22).

The principal change in Chapter 2, Reliability/Precision and Errors of Measurement, is signaled by the addition of the word “precision” in the title. Standard 2.0 expresses the primary intent of the chapter: “Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use” (2014 Standards, p. 42). Particular emphasis for credentialing should be directed to part of the narrative for standard 2.0: “A higher degree of reliability/precision is required for score uses that have more significant consequences for test takers” (p. 42). This is followed by eight clusters focusing on measures of reliability/precision, errors of measurement, decision consistency and documentation.

Chapter 3, Fairness in Testing, represented some of the most significant organizational changes in the 2014 Standards with consolidation of three 1999 chapters into one that now appears in Part I, Foundations. Camara and Tippins (2015) cited three primary reasons for these changes:

1. “All test takers have a right to fair and accessible assessments.”
2. “All individuals should be able to understand and respond to assessments without their performance being influenced by construct irrelevant characteristics of the test.”
3. “All test takers should have an unobstructed opportunity to reduce construct-irrelevant barriers.”

Standard 3.0 consolidates these objectives:

### TABLE 2. Chapter 3 – Standards for Fairness in Testing

<table>
<thead>
<tr>
<th>2014</th>
<th>1999</th>
<th>Comment (Removed 1999 standards 7.3-7.4, 7.9-7.10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0</td>
<td>New</td>
<td>Test development should minimize construct-irrelevant variance</td>
</tr>
<tr>
<td><strong>Cluster 1</strong></td>
<td><strong>New</strong></td>
<td><strong>Test Design, Development, Administration and Scoring Procedures that Minimize Barriers to Valid Score Interpretations for the Widest Possible Range of Individuals and Relevant Subgroups</strong></td>
</tr>
<tr>
<td>3.1</td>
<td>New</td>
<td>Test development should promote valid interpretations for all individuals</td>
</tr>
<tr>
<td>3.2</td>
<td>7.7</td>
<td>Broadens construct-irrelevant characteristics (culture, cognitive, physical, communication)</td>
</tr>
<tr>
<td>3.3</td>
<td>New</td>
<td>Test development should include subgroups in preliminary studies</td>
</tr>
<tr>
<td>3.4</td>
<td>7.12</td>
<td>Test takers receive comparable treatment in administration and scoring</td>
</tr>
<tr>
<td>3.5</td>
<td>7.2</td>
<td>Test development must document how construct-irrelevant barriers are removed from examinations</td>
</tr>
<tr>
<td><strong>Cluster 2</strong></td>
<td><strong>New</strong></td>
<td><strong>Validity of Test Score Interpretations for Intended Uses for the Intended Examinee Population</strong></td>
</tr>
<tr>
<td>3.6</td>
<td>7.1</td>
<td>Significant difference in subgroup scores may be defined by applicable laws; consider the appropriateness of test content for individuals with differences such as disabilities or limited language proficiency.</td>
</tr>
<tr>
<td>3.7</td>
<td>7.6</td>
<td>Use of regression analysis moved to comment</td>
</tr>
<tr>
<td>3.8</td>
<td>New</td>
<td>Evidence of validity of interpretations in scoring for relevant subgroups</td>
</tr>
<tr>
<td><strong>Cluster 3</strong></td>
<td><strong>New</strong></td>
<td><strong>Accommodations to Remove Construct-Irrelevant Barriers and Support Valid Interpretations of Scores for Their Intended Uses</strong></td>
</tr>
<tr>
<td>3.9-3.14</td>
<td>New</td>
<td>Validity for accommodations, evaluate changes, languages, interpreters</td>
</tr>
<tr>
<td><strong>Cluster 4</strong></td>
<td><strong>New</strong></td>
<td><strong>Safeguards Against Inappropriate Score Interpretations for Intended Uses</strong></td>
</tr>
<tr>
<td>3.15</td>
<td>New</td>
<td>Evidence supporting claims about applicability of test for subgroups</td>
</tr>
<tr>
<td>3.16</td>
<td>7.2</td>
<td>Same</td>
</tr>
<tr>
<td>3.17</td>
<td>7.8</td>
<td>Broadens construct-irrelevant characteristic (culture, cognitive, physical, communication)</td>
</tr>
<tr>
<td>3.18</td>
<td>7.5</td>
<td>Multiple sources narrowed for diagnostic/placement into opportunities to learn</td>
</tr>
<tr>
<td>3.19</td>
<td>New</td>
<td>Provide examinees opportunity to learn when same organization provides both curriculum and test</td>
</tr>
<tr>
<td>3.20</td>
<td>7.11</td>
<td>Same</td>
</tr>
</tbody>
</table>

Adapted from: Camara & Tippins (2015).
“All steps in the testing process … should be designed in a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population” (2014 Standards, p. 63).

Table 2 summarizes the Fairness standards in the 2014 Standards. Fourteen of the 20 standards are new or reconfigured.

**Part II Operations**

Part II combines Chapters 4 through 9 of the 2014 Standards to address the nuts and bolts of test development, scoring, reporting and documentation, then closes with rights and responsibilities of test takers and test users. Table 3 shows the overarching themes and clusters in Chapters 4 and 5. Credentialing programs may benefit from reaffirming how they address each cluster of these standards and updating any documentation that may help candidates and users of their results make better decisions based on their data and test results.

Two standards in Chapter 4 warrant special attention from credentialing programs.

- **Standard 4.3** reflects the increased attention to technology issues:
  “Test developers should document the rationale and supporting evidence for the administration, scoring and reporting rules used in computer-adaptive, multistate-adaptive, or other tests delivered using computer algorithms to select items. This documentation should include procedures used in selecting items or sets of items for administration, in determining the starting point and termination conditions for the test, in scoring the test and in controlling item exposure” (2014 Standards, p. 86).

- **Standard 4.16** is addressed by many programs, but some may need to expand resources to help candidates prepare for their tests based on the expansion of this standard:

> “The instructions presented to test takers should contain sufficient detail so that the test takers can respond to a task in the manner that the test developer intended. When appropriate, sample materials, practice or sample questions, criteria for scoring, and a representative item identified with each item format or major area in the test’s classification or domain should be provided to the test takers prior to the administration of the test, or should be included in the testing material as part of the standard administration instructions” (2014 Standards, p. 90).

An essential element in meeting Standard 4.16 is timing candidates’ access to sample materials and practice questions so that each test taker has the opportunity to learn, practice and understand the skills that will be assessed by the test. For example, a test may be designed to measure each candidate’s...
ability to use particular types of information or references that are part of on-the-job performance. Credentialing program managers should evaluate when access to these materials would be most beneficial for test takers as they prepare to take an examination. The more complex these resources and materials, the greater the importance of assuring that candidates have opportunities to understand how to use them BEFORE the date of the test.

Cluster 4 in Chapter 4 also benefits from the improved clarity the Standards provide for test revisions. Standard 4.24 gains added value in its direction that while a test does not require revision “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” (2014 Standards, p. 93). While this appeared as Standard 3.25 in 1999, its importance is easier to recognize in the 2014 edition. Standard 4.25 in the 2014 edition also gains added clarity as part of this section: “When tests are revised, users should be informed of the changes to the specifications, of any adjustments made to the score scale, and of the degree of comparability of scores from the original and revised tests” (p. 93).

Standard 5.16 confirms the Joint Committee’s effectiveness in updating the 2014 Standards to reflect the impact of current and future technology in scoring:

“When test scores are based on model-based psychometric procedures, such as those used in computerized adaptive or multistage testing, documentation should be provided to indicate that the scores have comparable meaning over alternate sets of test items” (2014 Standards, p. 106).

Similarly, Cluster 4 on cut scores in Chapter 5 adds clarity, even for the standards that did not change from the 1999 version. These standards specify documenting the rationale and procedures for establishing and validating cut scores.

Table 4 summarizes Chapters 6 and 7 by citing their overarching objectives and clusters. Most of these standards in Chapter 6 have not changed since the 1999 edition, but the added clarity from organization of the 2014 Standards should make each chapter easier to follow. The first seven standards address administration, followed by two standards that address scoring protocols and quality control, then a third cluster of seven standards that focus on score information and interpretation.

Similarly, Chapter 7 addresses documentation requirements for tests in four clusters focusing on appropriate use, test development, administration and scoring, and timeliness in delivering test documents. The standards have changed very little, but better organization should simplify and improve how they will be applied.

Table 5 presents the organization of chapters 8 and 9. Standard 8.1 reinforces the importance that all candidates have access to information about test content and purposes prior to testing, that it is available to all test takers, and that it is “available free of charge and in accessible formats” (2014 Standards, p. 106).
Cluster 2 emphasizes confidentiality and unauthorized disclosure and use of candidate scores. Cluster 3 addresses fair and accurate score reports.

Two standards in Cluster 4 of Chapter 8 deserve special attention from all credentialing programs.

1. **Standard 8.9:** “Test takers should be made aware that having someone else take the test for them, disclosing confidential test material, or engaging in any other form of cheating is unacceptable and that such behavior may result in sanctions” (2014 Standards, p. 136).

2. **Standard 8.12:** “In educational and credentialing testing programs, a test taker is entitled to fair treatment and a reasonable resolution process ... regarding charges associated with testing irregularities, or challenges issued by the test taker regarding accuracies of the scoring or scoring key. Test takers are entitled to be informed of any available means of recourse” (2014 Standards, p. 137).

Chapter 9 focuses on test users’ responsibilities, primarily to candidates. Standard 9.5 addresses the need to remain alert to possible scoring errors and to take corrective steps promptly when errors are suspected. Standard 9.11 addresses the importance of assuring that all candidates are measured appropriately in populations that include individuals with limited proficiency in the language of the test when translations or test adaptations are not available. Standards 9.17 and 9.18 address opportunities to appeal challenged scores and to retake an examination.

Cluster 3 in Chapter 9 addresses security, copyrights, and intellectual property rights, including “tests that are administered via electronic devices” (Standard 9.22, page 148). Standard 9.23 closes this chapter by specifying that “test users should remind all test takers, including those taking electronically administered tests, and others who have access to test materials that copyright policies and regulations may prohibit the disclosure of test items without specific authorization” (2014 Standards, p. 148).

**Part III   Testing Applications**

Just one of the four chapters in Part III is likely to have major significance for most credentialing programs. Chapters 10 (Psychological Testing), 12 (Educational Testing and Assessment), and 13 (Program Evaluation, Policy Studies and Accountability) address the needs of other audiences for the Standards.
Chapter 11, Workplace Testing and Credentialing, differentiates better than earlier editions between employment for selection and testing in professional and occupational credentialing. It reorganizes the standards into three clusters:

- Standards that apply to both employment and credentialing testing
- Standards that apply to employment testing
- Standards that apply to credentialing testing

Table 6 highlights those that are likely to require the greatest attention from most credentialing programs.

Cluster 1 in Chapter 11 primarily addresses the need to define content and the purpose of a test clearly and link it to “the job or professional/occupational requirements” (2014 Standards, p. 178). These standards reflect the need to support interpretations of test and non-test information used to make decisions with appropriate validity evidence. These requirements are common to both employment screening and credentialing.

Six standards (11.5 - 11.9 and 11.12) in Cluster 2 focus on applications of predictor-criterion relationships in employment testing. Standards 11.10 and 11.12 address alternate forms of validity evidence in employment testing. Generally these have limited applicability to most credentialing examinations.

Cluster 3, Standards on Credentialing, reinforces the need to link credentialing tests to “the importance of the content for credential-worthy performance in an occupation or profession” (Standard 11.13, p. 181). Standards 11.14 and 11.15 address the need for evidence that supports decision consistency and procedures for combining scores, respectively. Standard 11.16 underscores the importance of linking test content to “credential-worthy performance” and the prohibition against adjusting scores “to control the number or proportion of persons passing the test” (2014 Standards, p. 182).

A Final Overarching Theme: Test Security and Integrity

A final guideline of this “user’s guide” for the 2014 Standards documents its consistent emphasis on test security and integrity. This starts with how the 1999 and 2014 Standards define test security in their respective glossaries. Narrowing and tightening this definition to focus on test content and the validity of intended uses of a test reflects changes both in technology and in intellectual property issues since the late 1990s. The shorter definition actually strengthens the concept to better reflect present practices and applications to technology in testing.

One example is illustrated by Circular 64 from the U.S. Copyright Office, which specifies procedures to copyright item
pools and test forms for secure tests. The procedure involves in-office review to verify that the copyright is appropriate but places no actual test content into copyright records. The filed version blocks out secure test content that has been verified in a meeting with a U.S. Copyright Office employee. These procedures are highly recommended to protect the test security and integrity of all secure examinations. It also may help strengthen enforcement efforts if copyright violations of secure test materials do occur.

Table 8 cites references to test security and integrity throughout the 2014 Standards. These citations include 12 of the 13 chapters in the new Standards. Introductory discussions mention test security and integrity issues in nine chapters. These concepts appear in 10 standards and in the comments for seven more. Thus test security and integrity are one of the primary themes addressed throughout the 2014 Standards.

### TABLE 7. 1999 and 2014 Definitions of Test Security

<table>
<thead>
<tr>
<th>1999 Definition of Test Security</th>
<th>2014 Definition of Test Security</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td>Protection of the content of a test from unauthorized release or use, to protect the integrity of the test scores so they are valid for their intended use.</td>
</tr>
</tbody>
</table>

Source: Camara & Tippins (2015).

### References


### TABLE 8. Test Security and Integrity in the 2014 Standards

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 2</td>
<td>2.16 C²</td>
<td>51, 3.4, 3.20 C²</td>
<td>Intro 3 80-81, 83</td>
<td>Intro 3 97</td>
<td>Intro 3 111,113, 6.6, 6.7, 6.14, 6.16 C²</td>
<td>Intro 3 124, 7.9, 7.12 C²</td>
<td>Intro 3 131-132, 8.2 C², 8.9, 8.11</td>
<td>9.0, 9.16, 9.21</td>
<td>Intro 3 162, 10.18 C²</td>
<td>Intro 3 170-171, 177</td>
<td>Intro 3 188, 12.16</td>
<td>13.8 C²</td>
</tr>
</tbody>
</table>

Source: Camara & Tippins (2015).

1 Standards are listed by number.
2 C indicates the reference is in comments.
3 Intro refers to page numbers in the opening discussion in each chapter.
A Review of the Newly Adopted NCCA Standards

BRIAN D. BONTEMPO, Ph.D.
principal consultant, Mountain Measurement, Inc.

Brian served on the NCCA Standards Revision Steering Committee and is a Psychometric Reviewer for the NCCA. The opinions expressed within are those of the author and do not reflect the views of the National Commission for Certifying Agencies or the Institute for Credentialing Excellence.

On November 26, 2014, the latest revision of the Standards for the Accreditation of Certification Programs was adopted by the National Commission for Certifying Agencies (NCCA). Although the NCCA Standards were designed for certification programs, they are relevant to licensure programs as well. This article will provide a high-level overview of the NCCA Standards, explain the relevancy of the NCCA Standards for licensure programs, describe the NCCA Standards revision process, and identify some of the changes in this revision of the NCCA Standards.

Overview

The purpose of the NCCA Standards is to delineate the attributes of a quality certification program. The NCCA Standards “address the structure and governance of the certifying agency, the characteristics of the certification program, the information required to be available to applicants, certificants, and the public, and the recertification initiatives of the certifying agency” (p. 1). The major themes of the NCCA Standards are resources, autonomy, conflicts of interest, representativeness, transparency, accountability, confidentiality, security, psychometric soundness, quality assurance, and maintenance of currency with practice. All of these are issues that licensure and certification share.

The NCCA uses the NCCA Standards to evaluate certification programs that apply for NCCA accreditation. It is important to note that the NCCA accredits certification programs (e.g., Certified Medical Assistant) not certification organizations (e.g., American Association of Medical Assistants) or certification examinations (e.g., National Registry of EMTs Paramedic Certification Cognitive Examination). Currently, over 300 programs have been accredited by NCCA.

The NCCA Standards are organized as a collection of standards each of which has a description, one or more essential elements, and commentary intended to communicate expectations and help programs interpret the standard. When used for accreditation purposes, the NCCA Standards are minimum standards, meaning that certification programs must meet each and every standard and its essential elements to be granted accreditation. Programs cannot waive or otherwise opt out of compliance with any of the standards. In this regard, they are quite different than the aspirational Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological
Association, & National Council on Measurement in Education, 2014), which explicitly state, “Evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every standard in this document” (p. 7).

Relevance of the NCCA Standards to Licensure

Although NCCA accreditation is voluntary, some regulatory bodies require practitioners of specific professions to have their skills certified by an accredited certification program. For example, OSHA regulation 29 CFR 1926 Subpart CC, released August 9, 2010, requires crane operators involved in construction to be certified by an accredited certification provider. Similarly, California CCR, Title 9, Section 13035(c) requires addiction counselors to be certified by an NCCA accredited certification program. The Advanced Practice RN (APRN) Consensus Model (2008), which was developed by the vast majority of nursing credentialing programs for the purpose of assisting state boards of nursing with the licensure, accreditation, certification, and education of APRNs, requires all APRN certification programs to be accredited by the Accreditation Board for Specialty Nurse Certification (formerly the American Board of Nursing Specialties) or the NCCA. This model, which impacts over 250,000 APRN practitioners, has been adopted by many states and is continuing to grow in its scope and reach. Although much of this legislation allows certification programs to choose their accreditation, the NCCA accreditation is arguably the most popular accreditation available. These examples demonstrate that application of the NCCA Standards has extended to the practice of regulation and further justifies their applicability to licensure.

The NCCA Standards Revision Process

The initial version of the NCCA Standards, which was then called the Standards for Accreditation of National Certification Organizations, was developed and adopted by section between 1977 and 1982. The initial version contained 46 statements that were arranged into nine general topics. Although the content did not change, the name of the NCCA Standards changed to the current name in 1987 as a result of the complete reorganization of the corporate structure. The NCCA Standards began its first revision process in 1999. This revision changed the structure to contain stated standards, required essential elements, and explanatory commentary, identical to the structure still in use today. This revision was officially adopted in 2001 and made effective in 2002. A minor content revision occurred in 2004, which clarified the requirements of the public member and the standard pertaining to test administration.

The most recent revision to the NCCA Standards began in 2013 with a call for volunteers. During the 18-month effort, a steering committee and three task forces comprised of over 50 volunteer certification industry leaders met regularly to draft and revise the NCCA Standards. The proposed NCCA Standards were submitted for public comment on September 6, 2014. Based on public comment, the NCCA Standards were modified slightly and presented for vote on October 24, 2014 to the 131 agencies with accredited programs. Of the 87 agencies that submitted a vote, 86% were in favor of adopting the new NCCA Standards, which were officially approved on November 26, 2014.

The new revision of the NCCA Standards goes into effect January 1, 2016. All NCCA accredited certification programs will be expected to comply with the new revision of the NCCA Standards at that time. New and renewing applicants will demonstrate compliance by completing the new version of the NCCA application, which should be released later this year. Programs that were accredited by NCCA before 2016 will not be required to submit any additional evidence of their compliance beyond the information provided each year to NCCA as part of their annual report.

The preamble of the NCCA Standards states that the revision process was guided by a few principles that include relevancy, currency, consistency, and distinctiveness. In addition, special attention was made during the revision process to involve individuals employed by corporate certification programs (e.g., Microsoft Certification Program) and to address the relevancy of the NCCA Standards for corporate certification programs. Corporate certification programs are different than more traditional certification programs because they are offered by an organization whose primary purpose is generally not certification, regulation, and/or licensure. Often times, the purpose of these programs is to assist hiring managers rather than protect the public. Corporate certification programs are also different from employment testing programs for several reasons, the most salient of which is that a credential is granted that has value to employers other than the corporation offering the certification program.

The 2013-2014 revision process used the verbiage of the existing NCCA Standards as a starting point. In other words, a zero-order revision was not pursued. However, the Steering Committee evaluated alternative ways in which to convey each of the standards. Ultimately, the Committee decided to maintain the current format, which provides a description of the standard, its corresponding essential elements and supplemental commentary.
Technology has changed greatly since the last revision of the NCCA Standards was adopted in 2004. At that time, certification programs submitted their applications to NCCA in paper format or electronically via a CD-ROM. Each application contained hundreds, if not thousands, of pages of documentation assembled into what became known as an NCCA application binder. In 2013, NCCA began requiring programs to submit their application for accreditation online through a website that requires applicants to address each essential element of the NCCA Standards separately by providing a narrative response and relevant supporting documentation.

It is logical to conclude that the change in submission format encouraged the NCCA, the NCCA Standards Revision Steering Committee and certification programs to think about the NCCA Standards more analytically. This may explain why, despite only a small increase in the number of standards, the latest revision of the NCCA Standards contains 87 essential elements, an increase of 40 from the prior revision, which contained 47. Given this increase, some may project that NCCA applications will take more time to complete. On the other hand, the essential elements of this revision are intended to more clearly and logically delineate the components of a certification program. As a result, some programs may actually save time when completing their application.

**Changes to the 2016 NCCA Standards**

From a big picture perspective, there were not any major changes to the latest revision of the NCCA Standards. The focus of the revision effort was to clarify the wording and expectations of some standards and to expand the applicability of the NCCA Standards to the growing number of complex certification programs that require candidates to pass more than one certification examination to be credentialed. In addition, effort was made to update the NCCA Standards to reflect changes in the practice of test security. Along the way, an additional standard was born that covers error handling and quality control (Standard 23).

This revision of the NCCA Standards maintains the level of quality expected by accredited programs. In other words, certification programs are not being asked, for example, to make their programs more autonomous or their certification examinations more reliable. However, some standards were added or expanded, which will require certification programs to provide additional information to demonstrate their compliance. For example, Standard 16 Essential Element A requires programs to submit an item development plan, something that was not required under the previous revision of the NCCA Standards.

Although these additions will require certification programs to submit more information, accredited programs should already have the required policies and procedures in place. Therefore, these additions should require nothing more of applicants than submitting additional documentation along with their application. Nonetheless, some programs may find it necessary to spend additional time documenting their activities, policies, and procedures.

The NCCA Standards now contain 24 standards. There are five standards that pertain to purpose, governance, and resources, nine standards that outline policies and procedures, nine standards that apply to assessment instruments, and one standard that pertains to the maintenance of accreditation.

One small but important change was to the first standard, which now clearly excludes programs that only certify the competency of individuals to practice a skill. Although skill is not explicitly defined by NCCA, one can infer that a skill is a learned ability to carry out a task. Tasks are much smaller in their scope than a job role or occupation, and, generally speaking, individuals who possess a skill may use that skill in a wide variety of job roles. In essence, skills are the building blocks of a job role, profession, occupation, or specialty area. Although the assessment of a skill can be complex, it is out of the scope of the NCCA Standards.

In the governance section, small modifications were made to the documentation pertaining to the involvement of public members and relevant stakeholders. In addition to having a voting public member on the certification board, programs must now document how the public interest is “routinely represented and protected” (Standard 2: p. 4). In addition, programs must “identify their stakeholders and provide an ongoing mechanism to solicit their input” (Standard 2: p. 4). It could be argued that the new NCCA Standards require more involvement rather than simply more documentation. Either way, this modification is a win for protecting the interests of external stakeholders.

The text explaining the relationship between the training and certification functions of a certification organization was clarified by delineating it as a new standard entitled “Education, Training and Certification” (Standard 3: p. 6). Although the wording has changed from the previous revision of the NCCA Standards, the intent of this standard is largely the same: to maintain an appropriate firewall between
A REVIEW OF THE NEWLY ADOPTED NCCA STANDARDS

organizational functions. Specifically, the first part of this standard requires programs to maintain impartiality between the education and certification divisions of a certification organization. The second aspect provides a much-needed list of dos and don'ts for organizations that offer and require education/training of individuals pursuing certification. This list provides much-needed clarity to what has been a confusing standard for some time.

In turning to resources, the NCCA Standards now explicitly enable certification programs to be financially supported by another entity as long as there is a formal, written agreement in place obligating the sponsoring organization to support the certification program. With respect to human resources, the NCCA Standards now require programs to provide evidence of appropriate oversight and monitoring of personnel performing certification activities. This is an important modification from an accountability perspective. Programs can no longer defer responsibility for the actions taken by a testing vendor. If a program does not have qualified resources in-house by which to monitor its testing vendors, it may behoove the program to periodically have the services reviewed or audited by a third party.

The standards that pertain to a program's policies and procedures were also reorganized and clarified. Evidence of a policy and procedure describing the retesting of failing candidates is now required (Standards 6, 7). Programs should already have this policy in place, although some may need to add this documentation to their candidate handbook or equivalent. In addition, the NCCA Standards now prohibit a program from unreasonably limiting access to certification (Standard 7). This standard will prevent programs from requiring certificants to pay expensive fees (e.g., membership fees) that support organizational activities separate from certification and recertification.

Some of the essential elements of the policies and procedures section now require programs to provide their rationale for the policies and procedures. Although it is beyond the scope and capacity of NCCA to thoroughly evaluate the rationale provided by a program, these additions will help prevent organizations from capriciously impacting their candidates.

The wording and intent of the essential element relating to the use and acceptance of alternative certification examinations was changed. Programs must now demonstrate the content and empirical equivalence of another certification examination to accept it as an equivalent. Demonstrating content and empirical equivalence requires quantitative data from both examinations, which realistically limits the use of alternative examinations to certification programs that are collaborating with the organization sponsoring the alternative. Stated another way, this standard prevents a certification program from “kidnapping” the certificants of another program.

Two new standards were created about confidentiality (Standard 10) and conflicts of interest (Standard 11), which clearly document that which was implied by the previous NCCA Standards. Since confidentiality is of utmost importance to all testing organizations, the requirements of this standard serve to guide the way in which programs document their confidentiality procedures. The standard about conflicts of interest requires programs to have all relevant individuals sign conflict of interest statements, a step that some organizations may not have taken with all relevant personnel in the past.

The changes to the psychometric standards were more substantive. A new standard was created regarding the composition of subject matter expert panels (e.g., standard setting panel). This standard (Standard 13) requires programs to have panels that are inclusive and representative, which was only implied in the prior verbiage. Since the procedures needed to obtain representativeness will vary from profession to profession, programs should seek the input of an experienced psychometrician prior to assembling panels of subject matter experts.

The job analysis and standard setting standards (Standards 14 and 17) do a much better job of outlining the information that must be contained in the research report documenting the outcomes of any studies. Also, these standards contain an additional essential element that requires programs to conduct these activities frequently enough to maintain the currency of the program.

A new standard was created about examination specifications (Standard 15). This standard requires programs to provide more detail in their specifications than has been required previously. This includes the objective of the exam, the trait being measured (e.g., cognitive or psychomotor), the distribution of content, the types of items that are to be used, the item and examination refresh design, and the psychometric specifications such as the total number of test items, total administration time, and number of test forms.

The verbiage concerning test administration was transformed into its own standard (Standard 18) and more accurately reflects the objectives of a successful test administration. Stated in layperson terms, programs must ensure that
test administration is fair, confidential, and secure. Those interested in remote proctoring will find that the test administration standard neither permits nor prohibits this strategy. Rather, it identifies the aspects of a successful test administration. Although this is encouraging news for remote proctoring advocates, there is still a lack of third party research suggesting that remote proctoring is equivalent to or more secure than the face-to-face proctoring methods in use today. As a result, despite the new verbiage, programs using remote proctoring may still find it challenging to become accredited.

The scoring, reliability, and equivalency standards (Standards 19, 20 and 21, respectively) were reworded for clarity. In addition, an essential element was added to the scoring standard that requires programs to provide additional performance information to failing candidates, and the reliability standard now requires conjunctively (i.e., multiple-hurdle) scored examinations to have sufficient reliability in each of the independently scored sections. The equivalency standard now makes it explicit that programs must use empirical procedures to equate forms. The commentary indicates, “The use of standard-setting procedures in place of equating procedures is generally unacceptable” (Standard 21: p. 26). This is a marked change from the previous revision and may impact some programs.

A new standard was adopted that pertains to quality control and error handling (Standard 23). This standard requires programs to monitor activities and to take appropriate action when irregularities or program errors occur. From a test taker’s perspective, this may be the most significant change to the new revision of the NCCA Standards and should be welcomed by the public as a way of providing some assurance that programs are not turning a blind eye to the unfortunate technical errors that sometimes occur. Unfortunately for the public, this standard does not go far enough to ensure that test results will be corrected if an organization discovers an error.

Similar to the modifications made to other standards, the recertification standard (Standard 22) will require programs to provide additional documentation. Programs will need to document their rationale for the recertification requirements, as well as their process for verifying that certificants have met the recertification requirements.

The maintenance of certification standard (Standard 24) added an essential element that allows NCCA to conduct a program audit. The commentary indicates that this audit may be onsite, virtual, or through a third party. This new element will discourage programs from providing false information, thereby enhancing the quality and credibility of the NCCA accreditation. This also increases the value and motivation for individuals to file a complaint against a program with NCCA. Although NCCA is not likely to audit every program against which a complaint is filed, it is likely that NCCA will exercise its right to audit during 2016 and beyond.

Although the new NCCA Standards are longer and require programs to submit more documentation, they do not increase the level of quality required by certification programs. Nonetheless, the additional detail may reveal weaknesses in some programs, which may find it necessary to modify their policies and procedures in order to maintain their compliance with the NCCA Standards. Programs are advised to read the NCCA Standards carefully and to ask questions when they arise.

In reflecting on the changes as a whole, the NCCA Standards are more objective, which will make them easier for the certification programs and the NCCA to understand, implement, and enforce. The new NCCA Standards contain some new clauses, such as those in the governance and error-handling standards, that support the interests of test takers and the public. Other new clauses, such as those in the financial resources and use of alternative certification examinations standards, better protect the interests of the certification program. It will be interesting to see if the addition of the clause permitting NCCA to audit a program will increase the attractiveness of the NCCA Standards to legislators and regulatory bodies. Although the length of the new NCCA Standards may intimidate some, it is likely that most will find plenty of changes to their liking.

References


Subscore reporting: An updated reminder from the revised *Standards for Educational and Psychological Testing*

CORINA M. OWENS, Ph.D.
VINCENT LIMA
Professional Testing, Inc.

The *Standards for Educational and Psychological Testing* (APA, AERA, & NCME, 2014) (Standards) was recently revised and updated. This revision mainly has implications for educational measurement but does impact the world of certification and licensure testing. As Sackett (2014) indicates, the revision was “a modest updating rather than a major revision.” More specifically, the workplace testing and credentialing chapter contains a few changes with more of an emphasis on clarifying standards than on presenting new standards. However, these changes could impact how credentialing programs are setting policy.

As credentialing organizations begin to process what the revised version of the Standards means to them, it is important to point out exactly what changed within the workplace testing and credentialing chapter. The main change throughout the chapter is organizational. The chapter is now divided into three overarching areas: standards that apply to workplace testing, standards that apply to credentialing, and standards that overlap both areas. In addition, throughout the Standards there is reference to other chapters that could be useful in workplace testing or credentialing. This decision by the Joint Committee on the Standards encourages readers to remember to reference all applicable standards, not simply those within the workplace testing and credentialing chapter. The final change in the workplace testing and credentialing chapter focuses on addressing issues previously left out of the 1999 version of the Standards (AERA, APA, & NCME). Specifically, a standard was added that addresses the use of decision rules for multisection tests, and statements were added to clarify the use of subscore reporting.

As illustrated above, most of the workplace testing and credentialing chapter remained the same. However, statements about the need for reliable subscore reporting are of vast importance, as is the design of examination score reports. Subscore reporting has often been a topic of debate within the field of credentialing.

“To report or not to report?” is often the question. Examinees, or candidates as they are known in the field of credentialing, have a desire to understand their strengths and weaknesses as they relate to different sections of the larger examination. This desire comes from the high-stakes decisions made with regards to obtaining the credential or not.

The thought process in favor of subscore reporting is that this type of feedback can benefit candidates in improving their knowledge and scores and therefore help them to obtain...
Dear John:

Thank you for your participation in the Alphabet Exam. This is to confirm that your test score on March 10, 2015, was 67 percent. Unfortunately, your score was not within the passing range of 70 percent or higher.

Below you will find a list of the sections covered in the exam and how you performed in each one. These section scores are less reliable than your overall score since they are based on fewer questions. However, they can help guide you to topics on which you should focus your study if you decide to retake the exam. We strongly encourage you to do so. Experience shows that the majority of candidates who retest are successful. (All estimates of reliability and standard errors can be located at www.alphabetexam.com.)

<table>
<thead>
<tr>
<th>Section</th>
<th>Percent Correct*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section 1: ABC (16% of the exam)</td>
<td>75%</td>
</tr>
<tr>
<td>Section 2: DEF (9% of the exam)</td>
<td>92%</td>
</tr>
<tr>
<td>Section 3: GHI (11% of the exam)</td>
<td>64%</td>
</tr>
<tr>
<td>Section 4: JKL (16% of the exam)</td>
<td>8%</td>
</tr>
<tr>
<td>Section 5: MNO (22% of the exam)</td>
<td>0%</td>
</tr>
<tr>
<td>Section 6: PQR (9% of the exam)</td>
<td>78%</td>
</tr>
<tr>
<td>Section 7: STU (6% of the exam)</td>
<td>44%</td>
</tr>
<tr>
<td>Section 8: VW (3% of the exam)</td>
<td>70%</td>
</tr>
<tr>
<td>Section 9: XYZ (7% of the exam)</td>
<td>80%</td>
</tr>
</tbody>
</table>

*Out of a possible 100% in each section.

To arrange for a retest, please call 800-222-3333.

Sincerely,

Society of the Alphabet


to be interpreted, estimates of relevant indices of reliability/precision should be reported” (p. 43).

While the field of measurement cannot come to consensus on the most appropriate method of reporting subscore information, practitioners can agree that reporting unreliable information to candidates can do more harm than good. Sinharay, Haberman, and Puhon (2007) stress that just as inaccurate information at the total score level can lead to inaccurate classifications of pass/fail decisions, so too can inaccurate or unreliable information at the subscore level lead to incorrect remediation decisions. By the same token, Tate (2004) attests that validity and reliability are important not only for the total score of an examination, but also for any and all subscores reported in an effort to align remediation with actual weaknesses.

As illustrated above, the need to provide reliability/precision indices is a must for credentialing programs. Credentialing programs can provide reliability/precision information in a variety of ways. It has been suggested in the Standards that conditional standard errors be provided around all reportable information, both total scores and subscores. Along with or in place of the standard errors, confidence intervals can be calculated to provide a range of values somewhere within which the candidate’s true score lies.

However, even if subscore reliability is within acceptable limits, how should this information be presented to candidates in an effort to help them best utilize the information? As mentioned earlier, there is no consensus in the field as to what method works best. Credentialing programs can present information to candidates in a multitude of ways, and the best way to present score report information will depend on the audience of each credentialing program.

One method of score report presentation is to provide a text breakdown of pass/fail status and a numerical description
provided for some of the individual sections (for example, Section 8, which accounts for 3 percent of the exam) may be unacceptably low. A solution is to combine sections that can meaningfully be combined. The Standards in comments on Standard 2.3 acknowledge, “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” (p. 43).

Another method of score report presentation goes a step beyond a simple numerical breakdown of diagnostic feedback and provides qualifying descriptions of how the candidate performed. The number of levels described should depend on the precision of the outcome variable. For example, should the reliability of the outcome variable be high, then the credentialing program can feel more confident in providing more precise qualifying descriptions. An example of the qualifying description could be “Minimal,” “Basic,” and “Meets standard.” It is important to note that a description as to how the qualifying descriptors were created should be housed somewhere that is accessible to all candidates (See Figure 3).

In setting the performance levels for each descriptor, it is important to consider cases where a candidate fails the examination but receives diagnostic feedback that suggests he or she should have passed. Thus, if the sum of the scores required to earn the descriptor “Meets standard” in all areas is lower than the passing score, a candidate could “meet” the standard across the board but fail.

A final method of score report presentation compares the candidate’s performance in each section with the average performance of passing candidates. A candidate earning a score of 70 percent in a section may have done quite well (if the average score on that section is, say, 50 percent) or very poorly (if the average score on that

---

**FIGURE 2.** Notification of exam results: Numerical breakdown of diagnostic feedback with data visualization.

Dear John:

Thank you for your participation in the Alphabet Exam. This is to confirm that your test score on March 10, 2015, was 67 percent. Unfortunately, your score was not within the passing range of 70 percent or higher.

Below you will find a list of the sections covered in the exam and how you performed in each one. These section scores are less reliable than your overall score since they are based on fewer questions. However, they can help guide you to topics on which you should focus your study if you decide to retake the exam. We strongly encourage you to do so. Experience shows that the majority of candidates who retest are successful. (All estimates of reliability and standard errors can be located at www.alphabetexam.com.)

<table>
<thead>
<tr>
<th>Section</th>
<th>Percent Correct*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section 1: ABC (16% of the exam)</td>
<td>75%</td>
</tr>
<tr>
<td>Section 2: DEF (9% of the exam)</td>
<td>92%</td>
</tr>
<tr>
<td>Section 3: GHI (11% of the exam)</td>
<td>64%</td>
</tr>
<tr>
<td>Section 4: JKL (16% of the exam)</td>
<td>58%</td>
</tr>
<tr>
<td>Section 5: MNO (22% of the exam)</td>
<td>60%</td>
</tr>
<tr>
<td>Section 6: PQR (9% of the exam)</td>
<td>78%</td>
</tr>
<tr>
<td>Section 7: STU (6% of the exam)</td>
<td>44%</td>
</tr>
<tr>
<td>Section 8: VW (3% of the exam)</td>
<td>70%</td>
</tr>
<tr>
<td>Section 9: XYZ (7% of the exam)</td>
<td>80%</td>
</tr>
</tbody>
</table>

*Out of a possible 100% in each section.

Sincerely,
Society of the Alphabet

of diagnostic feedback for failing candidates. This method allows for a statement notifying the candidate whether they passed or failed the examination. Failing candidates would then be provided a more detailed description of their actual score, raw, percentage, or scale, and then a breakdown of performance within each content domain (See Figure 1). In addition, it is possible to use simple data visualization to enhance this kind of score report (See Figure 2).

The most likely objection to the reports shown in Figures 1 and 2 is that the reliability of the information

---

**FIGURE 2.** Notification of exam results: Numerical breakdown of diagnostic feedback with data visualization.
Dear Jane:

Thank you for your participation in the Alphabet Exam. This is to confirm that your test score on March 10, 2015, was 67 percent. Unfortunately, your score was not within the passing range of 70 percent or higher.

Below you will find a list of the sections covered in the exam and how you performed in each one. These section scores are less reliable than your overall score since they are based on fewer questions. However, they can help guide you to topics on which you should focus your study if you decide to retake the exam. We strongly encourage you to do so. Experience shows that the majority of candidates who retest are successful. (All estimates of reliability, standard errors, and explanations of level of mastery can be located at www.alphabetexam.com.)

<table>
<thead>
<tr>
<th>Section</th>
<th>Level of Mastery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section 1: ABC (16% of the exam)</td>
<td>BASIC</td>
</tr>
<tr>
<td>Section 2: DEF (9% of the exam)</td>
<td>MEETS STANDARD</td>
</tr>
<tr>
<td>Section 3: GHI (11% of the exam)</td>
<td>BASIC</td>
</tr>
<tr>
<td>Section 4: JKL (16% of the exam)</td>
<td>MINIMAL</td>
</tr>
<tr>
<td>Section 5: MNO (22% of the exam)</td>
<td>BASIC</td>
</tr>
<tr>
<td>Section 6: PQR (9% of the exam)</td>
<td>BASIC</td>
</tr>
<tr>
<td>Section 7: STU (6% of the exam)</td>
<td>MINIMAL</td>
</tr>
<tr>
<td>Section 8: VW (3% of the exam)</td>
<td>BASIC</td>
</tr>
<tr>
<td>Section 9: XYZ (7% of the exam)</td>
<td>BASIC</td>
</tr>
</tbody>
</table>

The following three diagnostic categories are used:

- Minimal: You have demonstrated limited evidence of knowledge in this area. This is an area in which you should focus your study efforts for future exams.
- Basic: You have demonstrated knowledge in this area. Further study may raise your overall score for the exam.
- Meets Standard: You have demonstrated ample knowledge in this area. Further study is unlikely to yield a noticeable improvement of your overall score.

The purpose of the Alphabet Exam is to measure the candidate’s level of knowledge compared to the criteria established for certification by the Society of the Alphabet. The results of the exam are not valid for selection purposes or for employment ranking within the alphabet industry because the test is not constructed to differentiate among candidates except at the passing score.

To arrange for a retest, please call 800-222-3333.

Sincerely,
Society of the Alphabet

Data visualization allows information about a candidate’s exam results to be communicated through graphic representation. In this example, the relative standing of the candidate is shown with arrows and shading; the height of each box represents the size of the section within the larger exam.
Dear John:

Thank you for your participation in the Alphabet Exam. This is to confirm that your test score on March 10, 2015, was 67 percent. Unfortunately, your score was not within the passing range of 70 percent or higher.

Below you will find a list of the sections covered in the exam and how you performed in each one. These section scores are less reliable than your overall score since they are based on fewer questions. However, they can help guide you to topics on which you should focus your study if you decide to retake the exam. We strongly encourage you to do so. Experience shows that the majority of candidates who retest are successful. (All estimates of reliability and standard errors can be located at www.alphabetexam.com.)

To arrange for a retest, please call 800-222-3333.

The Alphabet Exam measures each candidate’s level of knowledge compared to the criteria established for certification by the Society of the Alphabet. Every candidate who demonstrates that level of knowledge by achieving the passing score of 70 percent passes the examination. The section scores above are provided only as a guide to further study.

Sincerely,
Society of the Alphabet

The methods presented above only represent a few ways to summarize score report information for candidates. The revised Standards remind us that how we report score information to examinees is just as important as other aspects of exam development. While the Standards have provided guidelines on subscore reporting, they are simply a starting point in our effort to communicate with our candidates, and it is our responsibility to promote proper understanding of the score information we are providing.

References


This column features the “catch-up” inclusion of the 2013 item-writing and validation book by Haladyna and Rodriguez, an updated and expanded volume with substantial new content.

The breaking news in this issue is the publication of the 2014 edition of the Standards for Educational and Psychological Testing. The Standards are reviewed in a separate article by Rodgers in this issue; however, we do make note of the special Winter 2014 issue of Educational Measurement: Issues and Practice, which is devoted to comments on the process of development of the standards and the new document. Other topics based on recent publications include examination development, passing score studies, automated scoring of constructed response items, test subscores, and job task analysis.

What You Always Wanted to Know About Test Items

Haladyna’s book Developing and Validating Multiple-Choice Test Items (3rd ed, 2004) has been transformed into a volume that is more general in scope, with chapters on selected-response items (multiple-choice and beyond) plus constructed-response items. Transformation is not a term to be taken lightly, nor should it diminish previous accomplishments, going back to 1989 publications of research on item writing rules by Haladyna and Downing. Research on item development, as a “down in the weeds” applied activity has never received the attention of more glamorous technical subjects, but it is of great importance to the everyday lives of testing practitioners.

A comparison of the third edition of the previous title with the current volume reveals 140 more pages, updated references, and additional topics. The chapters on selected-response items (a more general term than multiple-choice test items) are well documented with current research and profusely illustrated with examples. The authors state that development and validation of constructed response items is not a well-researched area, but they do provide a recommended framework for development and evaluation as well as examples.

The book is divided into six parts and twenty chapters. The major sections are: I) A Foundation for Developing and Validating Test Items, II) Developing Selected-Response Test Items, III) Developing Constructed-Response Items, IV) Unique Applications for Selected-Response and Constructed-Response Formats, V) Validity Evidence Arising From Item Development and Item Response Validation, and VI) The Future of Item Development and Validation. Rather than attempt to summarize the content of the book, the following comments have been extracted as a stimulus to further learning.
“Despite the cognitive taxonomy’s considerable reputation, it has not been adequately validated….Studies of internal structure that reflects the six categories of the cognitive taxonomy have failed to yield evidence” (p. 33).

“Without any reservation, the three-option MC (multiple choice) is superior to the four- and five-option CMC (conventional multiple choice). Four- and five-option CMC should not be used, unless the options are logically based on common student errors, and an item analysis reveals that all distractors are working as intended” (p. 67).

“Given widespread support among testing experts, TF (true-false) is recommended for instructional testing with the caveat that it be done well. For standardized testing programs, we have other formats…that are more useful and have less negative research” (p. 71).

“The MTF (multiple true-false) format is an effective substitute for the complex MC. Because the MTF has inherently good characteristics for measuring knowledge and some skills, it should be more widely used” (p. 73).

“…innovative (computer) formats that reproduce cognitive demands already measured with the more conventional formats are hardly worth the effort” (p. 129).

“As the technology for AIG (automated item generation) continues to grow and improve, we will see the time when items will be generated by computer, probably for adaptive testing sessions…However, current methods seem limited to content that is decidedly quantitative” (p. 151).

For scoring an essay, “unless the intended inferences are focused on writing mechanics, response features such as handwriting legibility, sentence structure, grammar, punctuation, neatness, and spelling should be considered irrelevant” (p. 242).

“The SP (standardized patient) program is highly successful in training in the health professions and medicine especially where there is a doctor-patient interaction. The technique has research and development to support its validity” (p. 289-290).

“The portfolio is a time-honored device for obtaining a more comprehensive description of a candidate’s ability in a profession. However, it has threats to validity that should be recognized and studied. A portfolio is an advocate’s biased view of competence. The person who submits a portfolio does not submit evidence that runs contrary to the goal of certification. Also, as most portfolio scoring involves subjective rating, we have the same attendant threats to validity- rater inconsistency and rater bias…” (p. 296-297).

“Item development and validation continue to be a central aspect of test development. The scholarly effort to improve items is reaping rewards but there is much work ahead. The greatest needs are (a) unification of learning theory, (b) item-writing theory that leads to an item development technology, (c) more effective use of item formats, (d) validated guidelines for item development, and (e) continued growth and use of technologies that enable and assist item development and validation” (p. 415).


Plake, B.S. and Wise, L.L. What is the role and importance of the revised AERA, APA, NCME Standards for Educational and Psychological Testing? pp. 4-12.


Sackett, P.R. An employment testing and credentialing perspective on the Standards for Educational and Psychological Testing. pp. 22-34.

Ferrara, S. Formative assessment and test security: the revised Standards are mostly fine; our practices are not. pp. 25-28.


Zumbo, B.D. What role does, and should, the test standards play outside of the United States of America? pp. 31-33.


Sinha, S., Rijmen, F., Choi, S. and Dorans, N.J. The revised Standards and its role on research in educational measurement. p 36-38.

Maul, A. Justification is not truth and testing is not measurement: understanding the purpose and limitations of the Standards. pp. 39-41.

does provide complementary material on the background of the development of this edition of the Standards and their use, beginning with an editorial followed by a number of invited comments, intentionally selected to represent different perspectives and stakeholder issues.

The lead paper is written by Plake and Wise, co-chairs of the Management Committee responsible for revising the Standards, and they also provide a closing comment after the ten other papers. They begin with a history of the Standards and a description of the ten year long revision process. This includes the roles of the Management Committee and the Joint Committee, both of which had representation from the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education. A 2007 survey to identify needed changes in the 1999 Standards was conducted several years into the process, but the immense scope of the overall activity and attention to detail is captured in the fact that two rounds of public comments were solicited, and the second round in 2011 involved documenting responses to 4,000 individual comments. Plake and Wise (p. 9) indicate that “most of the standards and much of the background and discussion remain essentially the same” (from 1999). However, “the JC (Joint Committee) did make some significant changes to the content and format of the Standards. Chief among these are (a) reorganization of the chapters, (b) substantive changes to the chapters, particularly to the discussion of fairness, and (c) increased consistency across chapters” (p. 9).

Space limitations prohibit doing justice to the varied subjects of the ten response commentaries, but the titles listed above suggest the wide range of the discussion. At the risk of unduly calling attention to a minor point, the comments by Sinharay and colleagues and the response by Wise and Plake provide a context for the understanding of the purpose of the Standards. Sinharay and colleagues suggest that the Standards should include more specific guidance on topics such as acceptable thresholds for reliability of test scores and the value of subtests. They note that the Standards only state that “the distinctiveness and reliability of the subscores should be demonstrated (Standard 1.14) and that validity evidence supporting the subscores should be provided (Standard 1.15)” (36). Wise and Plake provide a reply to this perspective, stating, “The interpretation and use of subscores varies widely across educational, psychological, and employment. Even within educational uses, subscores might be used to make decisions about instruction for individual students or used to evaluate curriculum and instruction more generally. The Standards require ‘For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliability/precision should be reported.’ (Standard 2.3) It is then up to experts and users to decide whether precision is sufficient for particular interpretations…” (p. 42).

Test Item Statistics

The authors summarize the history of both classical test theory (CTT) and item response theory (IRT), indicating the advantages of IRT when assumptions of the IRT model are met. If the data are unidimensional and the test items fit the IRT model, the IRT model provides sample-independent results with scoring on an interval scale. The study was conducted using an examination administered in Turkey, the Turkish subtest of the 2003 High School Entrance Examination (HSEE). Data were obtained from 1250 candidates on the 25 items on the subtest. Factor analysis was performed on the data to determine if items on the subtest met the assumption of unidimensionality. As a result of this analysis, two items were dropped from the test. One-, two- and three-parameter IRT models were all reviewed, and it was determined that the three-parameter model best fit the data. A subsequent correlation analysis looked at the relationship between classical item difficulty (p-value) and calibrated item difficulty for each of the three IRT models. Additional correlations were calculated for CTT-IRT item discrimination statistics for the two- and three-parameter models. (The one-parameter model assumes that all items are equally discriminating.) Correlations for item difficulty (CTT-IRT) were very high for all IRT models, with the three-parameter model being lowest at .939 and the one-parameter model being essentially a straight line (.998). Discrimination correlations (CTT-IRT) were very high for the two-parameter model (.966) but substantially lower for the three-parameter model (.525). The authors were impressed by the similarity of CTT statistics and the one- and two-parameter models; however, they recognized that although the sample of examinees was relatively large, only 25 test items were included in the study.

Developing and Validating an Examination

This article describes the development of the Certified Management Accountant (CMA) examination: the content
outline, development of examination questions, and setting the passing score. The process is described as follows: The initial step is a job analysis survey and a literature review. The job analysis is used to develop the content specification outline and subsequently a blueprint that has a target number of items by content and three cognitive taxonomic levels. Questions are carefully constructed and reviewed. New questions are always pretested to obtain statistics. Passing scores are determined using a modified Angoff procedure.

Job Task Analysis Utilizing Multiple Focus Groups and Survey

Although this article is focused on the use of job task analysis (JTA) for development of translator certification examinations, it offers a perspective that should be of interest to anyone interested in JTA, as it provides rich detail on all aspects of the process beginning with the consideration of ISO standards and carrying through the development of an approved list of knowledge, skills, and abilities (KSAs).

The first section of the article covers the implications of the ISO 17000 series for accreditation, and this specific example would be of interest to other organizations that are considering pursuing ISO accreditation. For the job analysis, there is sometimes a need for clarification of the relative emphasis on a large number of tasks and the associated knowledge, skills, and abilities (KSAs) required. In this case a single task, translation, was taken as a given, and the focus of the JTA survey was on KSAs exclusively. Involvement of subject matter experts began with convening four focus groups: two primarily composed of freelance translators, one of government translators, and the fourth of translation project managers (p. 181-2). Subsequently, over 1400 responses were obtained in response to the survey that was developed. Thirty-six KSAs were prioritized based on a list of 13 knowledge areas, 13 skills, and 10 abilities. The survey also included six attributes to be ranked. The highest ranked attribute was “thorough, meticulous, and attentive to detail” (p. 185). Finally, results of the survey study were compared to other findings reported in the literature.

Angoff, Bookmark, a Hybrid Method, and Perspectives on External Validity
The Angoff method (Angoff, 1971, pp. 514-515) is based on either of two strategies. In one (provided in the original text), judges are asked to determine whether the *minimally acceptable person* would answer each question correctly. If the answer is “yes,” a score of 1 is calculated in the passing score. The recommended passing score is the sum of “1s” across all items of the test (the rater’s calculated passing score), with this final number correct value averaged over raters. Angoff offered an alternate method in a footnote in which the rater thinks of a probability that the minimally acceptable person will answer each item correctly. To get a passing score, probabilities are averaged over items and raters to get a percent correct passing score. The latter “footnote” method was far more popular for many years, but the method described in the text has been gaining ground. Modifications of the Angoff method are common, the most common being to provide data on level of item difficulty (percent correct), either initially or after the first round of ratings. These data provide a reality check on the ratings, although there is a danger of judges targeting their ratings to the level of difficulty of the item rather than the level of knowledge required by the minimally acceptable person.

The Bookmark method, in contrast, always uses item difficulty data and places items in difficulty order in a notebook. Judges place a bookmark between the pages that represent a set probability of correct response for the minimally competent candidate. Although conceptually there might be easy items that judges do not think that the minimally competent candidate would answer correctly or difficult items that the competent candidate would answer correctly, the underlying assumption is that competency can be identified by a position along a continuum of ordered item difficulty.


The authors note that the Angoff and Bookmark passing score methods are the most frequently used standard setting methodologies, yet both have been subject to criticism. The approach recommended in this article is intended to take advantage of the strengths of both methods. The Ordered Item Book (OIB) method is based on ordering of items from easiest to hardest in the review book. Judges are made aware of the ordering of the item book and then are asked to rate the items using an Angoff yes/no approach. The study compared the hybrid OIB method and a more-traditional Angoff yes/no method. The authors indicate that the “percent answering correct” version of the Angoff method is also compatible with the OIB approach. Advantages of the OIB method are cited, which focus on the fact that the items are in difficulty order, but the method allows raters to make
A study was conducted in which 13 panelists were divided into two groups (six and seven individuals) in order to compare the Angoff method (item book arranged in content order) with the OIB hybrid. Each method had two rounds. Specific information on item difficulty was shared in the second round, but the OIB group had the knowledge that their items were arranged in difficulty order in the first round. “The results show that members of the Angoff group, on average, made twice as many item-level changes” (p. 23). One conclusion of the study is that “as an initial look, the results from this study support the claim that results from the OIB Angoff method appear to align with results from another commonly used method (e.g., similar recommended cut scores to the Angoff method) with some potential advantage to the panelists (e.g., less challenging task, smaller amounts of variance among panelists)” (p. 25).


This article contains a review of perspectives on external validity evidence for passing scores and follows through with two examples of practical approaches “in settings where other national or international benchmarks may not be available to guide practitioners.” One approach is to use multiple methods for determining passing scores. The authors indicate that there is not a single superior method for determining passing scores. An illustrative study is provided in which two groups of 20 judges performed two standard setting studies: one a modified Angoff study and the other a Bookmark study. All judges participated in both methods, but the order of the passing score studies was reversed for the first and second groups. In both studies, there were two rounds, with p-values and impact data being provided on the second round. The authors conclude that “…there is moderate supporting validity evidence for both methodologies; more so for the modification of the Angoff method in this instance. The only substantial difference observed was within the external component when the respective impact of each methodology was compared to the expectations of the faculty working directly within the professional training programs. The results of this analysis suggest the recommended cut score from the Angoff methodology more closely aligned with the expected examinee classifications” (p. 64).

A second approach to external validity proposed is to gather comparison data for interpretation of impact of the passing score. The potential passing rates are compared with another similar assessment, or, if such an assessment is not available, comparison with results associated with a similar population. The authors use as an example an assessment being used internationally. In the absence of other national data to consider, they suggest comparing results with a different nation’s results, if there is comparable level of economic development, a similar educational accountability program, and similar performance on other international examinations.

Repeat Testing on the Same Exam Form


This study addressed the concern of repeat candidates being advantaged by taking the same examination form a second time. A study was conducted in which two forms of a 200-item certification examination were available. Normally repeat candidates were assigned a parallel form containing only 25% overlap with the previous year’s form. For this study, 43 of 338 repeat candidates were assigned to take the previous year’s form, and the majority of repeat candidates were assigned the parallel form. Repeat candidates in both groups had achievement gains, but there was not a difference between the two groups. Seeing exactly the same items did not have an incremental impact on the “identical form” group. Additional analyses were conducted on item response patterns and response times. Continued selection of the same incorrect answer option on 68% of the items led the authors to conclude that the repeat candidates tended to be misinformed rather than uninformed.

The study provides encouragement to organizations having small candidate volumes that may not be able to afford to offer a second form to repeaters; however, in the study there had been a tradition of having a different form available for repeaters. There might be a different outcome if repeat candidates were aware that only one form was available, and that they would likely be offered the same test form that they had previously received if required to repeat.

Validity Perspectives on American Board of Family Medicine Examinations


This article discusses the purpose of American Board of Family Medicine certification examinations and education of the public with respect to whether individuals have passed
the examination or not. They state, “ABFM examinations measure a physician’s fund of medical knowledge within the context of the clinical practice of the specialty of family medicine. The examinations do not measure other important aspects of family medicine such as one’s clinical or procedural skills, the ability to communicate with patients, professional skills and behaviors, the ability to practice within a system of care, and the ability to learn from the practice of family medicine to continuously improve patient care….It is critical that consumers understand that simply because a physician fails the Maintenance of Certification for Family Physicians (MC-FP) examination does not mean he or she is incapable of providing high-quality care or is incapable of becoming more knowledgeable about the important body of knowledge that defines the specialty of family medicine” (p. 430).


The Maintenance of Certification Examination for Family Physicians is an examination offered by the American Board of Family Medicine that “is intended to measure the single construct of clinical decision-making ability within the practice of family medicine” (p. 344). Confirmation that a single construct is measured would be evidence of validity for the use of the examination. The examination is based on the Rasch IRT model. The measurement model assumes a unidimensional test. In addition to reviewing the fit statistics for the examination, the investigators performed a factor analysis on standardized residual correlations. Based on the analysis, the conclusion was that the examination is “highly unidimensional from a psychometric perspective” (p. 34).

Food Safety Knowledge and Certification Status


This article describes the results of a study of the relationship of food safety certification and food safety knowledge conducted by the Centers for Disease Control. Several hundred restaurant managers and workers were interviewed in six different geographic locations around the country. “Analyses showed that certified managers and workers had greater food safety knowledge than non-certified managers and workers” (p. 835). A number of other trends in the data also were identified.

Correlation between In-training and Certification Examinations


The authors sought to determine the predictive value of the orthopedic-in-training examination (OITE) of the American Board of Orthopedic Surgeons and Canadian orthopedic surgery residents’ scores on the Royal College of Physicians and Surgeons of Canada (RCPSC) certification examination. For 118 candidates involved in the study, positive correlations were obtained between 2000-2003 OITE scores for years one and two of residency training and pass/fail status on the RCPSC examination. No candidates in year three failed the RCPSC examination, precluding this analysis. In addition to positive correlations between multiple-choice examination sections, positive correlations were also found between OITE scores and the oral section of the RCPSC examination. The study conclusion was that the OITE had the potential to predict the success of Canadian residents on the RCPSC examination, as the OITE and RCPSC examinations were structured during the time period covered by the study. (p. 262) Since 2009 changes have been made in the reporting of OITE scores, and an objective structured clinical examination has been added as a component of the RCPSC examination.

Impact of Interruptions in Online Testing


This article addresses a concern about scores on computer-based tests when the testing session is interrupted and the test must be restarted, either immediately or on a different day. A number of examples are cited for state-level testing programs. The goal of the study is to determine the impact of the interruptions on test scores. “Several methods, primarily based on propensity score matching, linear regression, and item response theory, were suggested to determine the overall impact of interruptions on the examinees’ score” (p. 419). Data were analyzed from both a simulation study and a statewide assessment. The authors conclude that “the interruptions did not have a significant overall impact on the scores for the ISTEP+ (Indiana Statewide Testing for Educational Progress-Plus) test” (p. 419). While the finding is good news for programs where group data is of primary interest, the scope of the conclusion does not apply to individual scores on certification and licensure tests. For these latter examinations, there is no way
of knowing for certain what an individual score would have been if there had been no interruption in the testing session.

Automated Scoring of Constructed Response Questions


Automated scoring systems for essays and other constructed response test items derive their value from the labor savings that they offer and derive their legitimacy from studies showing a high degree of agreement with the ratings of human raters. While accepting the value of automated scoring in general, this study addresses the question of whether candidates can “game” the system by padding their responses with material that is irrelevant to the construct being measured but is awarded credit by the scoring algorithm. The authors designed a study that included gaming strategies matched to a task, simulation of the gaming strategies, optimization of the strategies, and evaluation of the scoring methods (p. 38).

The experiment centered on a competition for scoring short answer questions. Three prize-winning scoring engines from the competition were assessed based on three different manipulations of responses: (1) adding to the length of the response by including multiple copies of the original response, (2) adding words from the item stimulus to the response, and (3) adding academic vocabulary to the response. None of these additions would impress a human rater, but they might enhance scores derived by automated systems. All three of the open source test engines investigated showed some increase in scores as a result of the gaming strategies, one showing an increase of almost a standard deviation.

Although this study was necessarily limited in scope, it certainly provides much food for thought. The discussion section of the paper includes the following comments.

“This paper has demonstrated the need for an evaluation metric that can assess how susceptible automated scoring systems are to construct-irrelevant strategies for achieving inflated scores. … One key lesson of this study is that simple gaming strategies can have non-negligible effects on the behavior of automated scoring engines. … A second lesson is that the susceptibility of a scoring engine to gaming strategies is not always readily apparent based on consideration of the logical structure of the scoring engine” (p. 43).

Reliability of Subscores and the Question of Added Value


As indicated in the title, the first article’s message is that in general it is not beneficial to score an item on more than one subtest. This perspective is presented in the context of a discussion of the demand for more information than is provided by an overall test score, the typically low reliability of scores on short subtests, and other factors such as the unidimensionality of many tests (making subscores less relevant) and the homogeneity of populations of many certification and licensure candidates (education and experience requirements making the candidates more similar, resulting in less variation in scores and lower measured reliability).

Reliability of a subtest is affected by the length of the test, but the length of the total test usually has practical limitations. Controlling the length of the total test and increasing subtest length by having items do double duty on multiple subtests does have some intuitive appeal, but the authors point out that this strategy can defeat the major purpose of the subtest, which is to provide different information from that provided on the total test score.

This perspective brings us to the equation that the authors offer to predict a subscore’s value. The proposed formula for calculating the value added ratio is

\[1.15 + 0.51 \times \text{subscore reliability} - 0.67 \times \left( \text{correlation of the subtest score with remaining total test after subtest items removed} \right) / \left( \text{square root of the product of the two reliabilities} \right)\]

If the value is greater than 1, then there is value to the subtest score. A worked example for two tests is provided.

References


Examination owners in the licensure and certification community go to great lengths to protect the security and integrity of the examination programs. The time and capital investments are significant and necessary to ensure examinees demonstrate the competence tested in the assessment instrument. All examination programs are subject to breaches, and the test prep industry is a lucrative business venture, creating incentives to develop and market materials that “better” prepare examinees to pass the relevant examination. Examination owners are encouraged to register protected materials with the United States Copyright Office to provide maximum legal protections and opportunities to pursue alleged wrongdoers. Not all legal disputes involve examination owners protecting licensure and certification examinations. Test prep companies also have intellectual property rights and may elect to litigate issues involving protected preparation materials. Consider the following.

Both parties to this litigation are examination preparation organizations that prepare customers for the examinations used as one criterion for licensure eligibility as contractors. The Louisiana Contractors Licensing Service, Inc. (Plaintiff) is a Louisiana corporation that “prepared contractors for...state licensing exams” that included the Louisiana state licensing exam. The American Contractors Exam Services, Inc. (Defendant) is a Tennessee corporation that also prepared contractors for state licensing exams, including the Louisiana exam. As part of its exam prep program, the Plaintiff used sample exam questions to assist its customers dating back to 1993. In January 2003, the Plaintiff registered its materials with the United States Copyright Office. Such registered materials consisted of item banks of sample questions covering content from the Plaintiff’s Business Law Study Guide, Residential Study Guide, and Building Construction Guide.

In its complaint filed in the United States District Court for the Middle District of Louisiana, the Plaintiff alleged that the Defendant began using Plaintiff’s protected copyrighted sample exam questions in the course of Defendant’s business. The Plaintiff alleged that it did not authorize the use of such sample exam items and Defendant’s infringing activities dated back at least to 2009. As a result, the Plaintiff alleged deliberate and willful copyright infringement in violation of federal law.

In response, the Defendant filed a motion for summary judgment. A motion for summary judgment argues that there are no genuine material issues of fact in dispute and the court can decide the legal issues without the necessity of a trial. The Plaintiff argued that there were 14 identical copyrighted questions used by the Defendant, except that the Defendant used fill in the blank questions and the Plaintiff utilized a multiple choice format. There was a procedural dispute as to when and whether the Plaintiff did or must turn over copies of items substantiating the infringing activities of the Defendant. The Plaintiff did eventually turn over copies of the exam items but not, according to the Defendant, until after the cut-off date for discovery.
On the merits, the Defendant argued that there were only 12 identical questions, that there was no copyright infringement because the Defendant used different style of questions, and that the questions addressed common issues, facts, and concepts that “any person preparing contractors for a licensing exam [would] go over.” The Defendant also claimed that any unauthorized use was de minimis and only amounted to 12 or 13 questions out of 1,083 copyrighted multiple choice sample questions. Finally, the Defendant argued that any unauthorized use as alleged is subject to the fair use doctrine. The fair use doctrine was codified into federal law in 1976 and basically allows copyrighted works to be used in limited settings such as education, scholarship, or research without violating the copyright protections.

In addressing the Defendant’s motion for summary judgment, the court first provided an overview of the standard by which such motions are analyzed. The party seeking summary judgment must show that there are no genuine issues of material fact in dispute and that such moving party is entitled to judgment as a matter of law. The opposing party has the opportunity to show that there are issues of material fact in dispute or that the moving party is not entitled to judgment as a matter of law. Next, the court ruled on the issue of whether the “late filed evidence” was able to be considered in the analysis of the motion. The Plaintiff argued that the Defendant did not timely produce copies of the test questions alleged to have been infringed and, thus, such evidence was inadmissible. In short, the court held that the late filed submissions were not necessarily the fault of the Plaintiff and, in addition, were not prejudicial to the Defendant. As a result, the court in making its determination considered the exam items submitted by the Plaintiff.

Allegations of copyright infringement require plaintiffs to establish two elements: (1) ownership of a valid copyright and, (2) copying of constituent elements of the work that are original. The second element requires an inquiry of whether the alleged infringer copied or “actually used the copyrighted material in his own work” as well as whether substantial similarity exists between the two works. In making a determination of use and similarity, “a layman must detect piracy ‘without any aid or suggestion or critical analysis by others’. ... The reaction of the public should be spontaneous and immediate.”

As noted by the court and propounded by the Defendant, a defense to copyright infringement is a claim that even if the materials were identical, the infringement is de minimis. “The de minimis doctrine provides that if unauthorized copying is sufficiently trivial, ‘the law will not impose legal consequences.’” This inquiry focuses on whether the infringement of the copyrighted materials is significant enough to constitute infringement. A defendant arguing this defense must show that the “copying of the protected material is so trivial as to fall below the quantitative threshold of substantial similarity”. According to the court, this principle of trivial copying has long been a part of copyright jurisprudence.

Citing previous case law, the court turned its attention to the magnitude of the alleged breach and noted that, taking the Plaintiff’s claims as accurate for purposes of summary judgment, 14 exam items out of an item bank of 1083 were compromised. Based on these uncontroverted facts, the court pointed out that 14 questions comprised 1.3% of the overall work, that being the item bank. The court compared this fact pattern to a previous case where a composer sued the Beastie Boys arguing copyright infringement over the use of a three-note lick used as a background loop in a portion of a song. In that case, the court found that “no reasonable juror could find the sampled portion of the composition to be a qualitatively or quantitatively significant portion of the composition as a whole,” as it comprised only six seconds of the song, or approximately two percent.

Similarly in the current case, the court referenced the de minimis nature of the infringing materials, 1.3%, relative to the whole, quantitatively insignificant. It also noted that no allegations were made by the Plaintiff that the infringing materials were any more significant than the other 1,069 items, indicating a de minimis qualitative use. Accordingly, the court found that the “simple, minimal, and insignificant” use of a portion of the entire copyrighted material justified granting the Defendant’s motion for summary judgment and dismissed the case.

Examination owners must be aware of the various defenses to allegations of copyright infringement, including fair use and de minimis infringement. Significant item harvesting schemes likely extend beyond such de minimis nature, but individual examinees sharing a few items may provide a basis for judicial analysis similar to this case. As a reminder, multiple other causes of action may exist against infringing participants, and exam owners are encouraged to explore any and all such options to protect the integrity of the examination program and the constituents served.

Reference

Introduction

CLEAR members, jurisdictions, boards, and other stakeholders are continually faced with new questions and practical issues about their examination programs. Numerous resources—including Resource Briefs, Frequently Asked Questions, and discussion forums—are provided on the CLEAR website to assist members in tackling such issues. At the annual conference, new information is shared through sessions and networking opportunities.

This column presents practical issues and topics from recent Ask the Experts conference sessions, where audience participants pose questions to a panel of testing experts. In this column, panelists present their perspectives on specific questions or issues raised at the CLEAR Annual Educational Conference.

These responses represent the views of the contributor, are specific to the situation, and offer general guidance. Each response represents the perspectives of the individual columnist and is not to be considered an endorsement by CLEAR. Psychometrics is a blend of science and art, and each situation is unique. The responses provide general background and guidance, which can be used to inform decisions with additional input from psychometricians to fully respond to your specific issue.

Readers are encouraged to submit questions for future columns and conferences to clearhq.org. Please enter “CER Perspectives on Testing” in the subject line.

Why is measuring continuing competence important? What are some methods and examples?

Response provided by Martha Jeanne Guernsey, Director, National Board of Medical Examiners (NBME)

Many health professionals and certifying organizations have come to recognize the importance of assessing continuing competence as a privilege and responsibility of self-regulation and in the interest of protecting the public. A recent article in the journal Contemporary Pediatrics highlighted, “No longer do professionals, parents, or the public believe that passing an exam once in a career is enough to demonstrate to all that a doctor will always be current” (Moyer, 2014, p.15). In fact, there is evidence that the public assumes (and expects) that licensed or certified health professionals are required to engage in additional training and demonstrations of competence throughout their careers. Results of a 2007 survey conducted by the American Association of Retired Persons (AARP) found that 90% of respondents believed that all licensed healthcare professionals should be required to periodically demonstrate competence (Stowell-Ritter, 2007, p.3).
Competence is not a static achievement, but an ongoing process of continuous learning and practice improvement. A professional’s competence must evolve along with advances in technology, science, knowledge, and other workplace and social changes. In the health professions, for example, one challenge is that the broad practice population often ages along with the physician, so the nature of the practice can change. For these reasons, researchers and organizations responsible for assuring safe and effective medical practice have recommended developing continuous competence requirements that include both education and assessment components (Coalition for Physician Accountability, 2014; Swankin, Lubuhn, & Morrison, 2006, p.iii; Swankin, 2010).

In medicine, education requirements should be targeted towards an individual’s practice rather than based on more generalized educational activities. The goal is to have a positive impact on practice improvement and better patient outcomes (Federation of State Medical Boards, 2014, p.5). Self-assessments, such as diagnostic testing or review of practice outcome data, can enhance achievement of this goal by identifying areas of strength and weakness that can be used as a basis for targeted subsequent educational activities. Such assessments may be repeated to evaluate whether knowledge is gained through such educational activities and to evaluate the impact of changes in practice. Periodic high-stakes assessments, such as a secure re-examination or review of actual practice, can protect the public by demonstrating that an individual continues to meet required minimum standards for a profession as its associated body of knowledge evolves over time. To this end, the International Standards Organization (ISO) 17024 standard 9.6.5 specifically recommends that a certification body should consider examination as a component of recertification requirements (ISO/IEC, 2012).

In the United States, the American Board of Medical Specialties (ABMS) Maintenance of Certification (MOC®) program includes both education and assessment. There are four measures: (1) licensure/professional standing, (2) lifelong learning (education targeted to areas identified for improvement) and self-assessment, (3) high-stakes examination, and (4) practice performance assessment. The assessment component is further elaborated to require measurement of both cognitive knowledge (“know how”), using self-assessment as well as a secure examination, and measurement of actual performance in practice (“show how”). These requirements are continuous, with defined activities that must be completed during every renewal cycle, and are based on six core competencies: (1) Practice-based Learning and Improvement, (2) Patient Care and Procedural Skills, (3) Systems-based Practice, (4) Medical Knowledge, (5) Interpersonal and Communication Skills, and (6) Professionalism (American Board of Medical Specialties [ABMS], 2015; ABMS, 2014, p.2).

In Canada, the Royal College MOC is based on earning 400 credits of continuing professional development activities every five years. The program addresses defined competencies, and credits are awarded for approved learning activities included in the “Framework of Continuing Professional Development Activities,” which covers (1) Group learning, (2) Self-learning, and (3) Assessment (Royal College of Physicians and Surgeons of Canada, 2013).

In the United Kingdom, the General Medical Council (GMC) provides an alternative continuing competence assessment framework for physician revalidation that incorporates a five-year cycle of activities. This process is comprised of performance appraisal and recommendations by a responsible officer based on the evaluation of a practitioner’s portfolio. The portfolio documents personal attributes, continuing professional development activities, practice review, feedback from colleagues and patients, and a review of any complaints and compliments that might have been filed concerning the practitioner (General Medical Council [GMC], 2013; GMC, 2012, pp. 3-12).

Regulatory and certification requirements drive continuing education, professional development, and assessment (Melnick, 2004, p. S47). The best methods for measuring competence are those that have the “right fit” for a particular profession. They must be cost effective and relevant to practice, assess the competencies essential for safe and effective practice, and enable appropriate inferences about the competencies being measured, thus ensuring the best possible outcomes commensurate with enhanced public trust.

References


What information/feedback should our jurisdiction provide to candidates?

Response prepared by Paul D. Naylor, Consultant, PDN Consulting

In many jurisdictions information for candidates depends on how well they performed on the examination.

For those passing, I recommend providing congratulations on having passed but no numerical score. It must be recognized that criterion-referenced examinations are not designed to rank individual candidates. Rather, the purpose is to indicate the candidate has demonstrated minimal competence for entry into the occupation. If policy or regulation requires that a score be given, it should be emphasized that the score should not be used for purposes such as advertisements or personal growth activities.

For failing candidates, I recommend providing their overall score and the required passing score. In order to help candidates prepare for another attempt they should be given feedback about their performance by domain, content area or subsection. It should be noted that this information, while helpful, may present some problems of its own. If your jurisdiction uses scaled scoring on reports, candidates (despite warnings to the contrary) may attempt to add up the subscores and find they do not equal the score reported. Additionally, it is sometimes difficult to explain scaled scores to the satisfaction of the failed candidate.

Another important caution in reporting is to understand that domain level information has low reliability. When reported to failed candidates, the section or subscore data should be accompanied by guidance on interpreting the scores. (This guidance is usually only understood by psychometricians and may be unhelpful to normal people.)

In many jurisdictions I have seen the use of bar graphs representing strength and weakness on reports to failing candidates. The graphs indicate gradations from low to high for each of the domains and show the relationship of one area to the other without giving numerical scores.

In addition to examination results, the failed candidate should be given information on the examination study guide, instructions for rescheduling, and the appeal process.

Another closely related issue is:

What do I tell the “nearly passed” candidate?

Response prepared by Paul D. Naylor, Consultant, PDN Consulting

This is a difficult area; the candidate is aware of being only one question away from passing and tends to ignore the other 28 items that were incorrectly answered. The following steps should be helpful in assuring the candidate that the exam was not put together in a haphazard manner.

- Investigate the candidate’s question; check the facts of the situation
- Make sure there were, in fact, no scoring errors
- Explain the job/practice analysis
• Describe the test plan and its development
• Explain the standard setting process
• Explain the test assembly process and score report information
• Review the score report, including all key features, with the candidate
• Explain scaled scores (if appropriate)

Some may ask if there were so-called trick questions on the exam. This is a good time to talk about the test plan development. Explain the item development process including test writer training, item pretesting, item review, and standard setting.

While these explanations will probably not completely satisfy candidates in this position, it is important to give the person your assurances that you are interested in how important this is to them and to their profession.