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, 108 Wind Haven Drive, Suite A, Nicholasville, KY 40356.

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
Retired, Applied Measurement Professionals, Inc.

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@proexam.org

Contents

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

COLUMNS

Abstracts and Updates ......................................................... 3
George T. Gray, Ed.D.

Legal Beat ............................................................................ 9
Dale J. Atkinson, Esq.

Perspectives on Testing......................................................... 12
Chuck Friedman, Ph.D.

ARTICLES

Item Writing: Methods for Generating a Greater Return on Investment ........................................ 16
Belinda Brunner, M.S.

Clarifying the Blurred Lines Between Credentialing and Education Exams ............................ 23
Chad W. Buckendahl, Ph.D.
For your fireside reading, we present the Winter 2016-17 issue of CLEAR Exam Review, volume XXVI, number 2. This issue offers three regular columns and two feature articles covering a variety of topics and perspectives.

George Gray’s Abstracts and Updates column focuses first on job analysis methodologies, reviewing recently published reports on job analyses in assorted professions. Next, test validity is spotlighted in a summary of articles presenting dueling conceptualizations of validity. Dr. Gray also synopsizes recent publications that address issues related to the use of multiple measures, equating and linking, public disclosure of test items, and the inclusion of performance assessments in a testing program. Finally, a correlational study of variables predicting success on the National Council Licensure Examination for Registered Nurses (NCLEX-RN) is reviewed.

Legal Beat by Dale Atkinson reviews the purpose and application of the Americans with Disabilities Act (ADA) in the context of licensure and certification testing and summarizes an ADA-related lawsuit initiated by a medical student against the National Board of Medical Examiners (NBME) and the Federation of State Medical Boards (FSMB). The column focuses mainly on issues related to discovery and disclosure of documentation the Plaintiff and Defendants requested of one another. The Court’s decisions help to clarify the types of information that might be reasonably requested in such a case and what factors affect judgments related to disclosure.

Perspectives on Testing, coordinated by Chuck Friedman, addresses questions raised by CLEAR members and conference attendees. Experts in psychometrics, exam development, and other aspects of testing provide responses. In this issue, experts answer questions regarding the reporting of sub-scores, the inclusion of easy items on an exam, and cost/benefit analysis of using innovative item types.

Next, Belinda Brunner presents alternative techniques for the generation of new test items in Item Writing: Methods for Gaining a Greater Return on Investment. Test security is a priority in licensure and certification, particularly for high-stakes examinations. Building and maintaining a large bank of viable test questions is an important line of defense, yet item generation can be quite expensive. This article describes tools and techniques for the efficient generation of effective test items.
Finally, in *Clarifying the Blurred Lines Between Credentialing and Education Exams*, Chad Buckendahl describes similarities and differences between testing in an educational setting and testing in a credentialing context. Three aspects of testing, in particular, warrant attention to these distinctive contexts: the relationship between assessment and curriculum/training; the focus of measurement (e.g., minimum competence versus a broader range of performance); and issues related to disability accommodations. Readers may find this article useful in helping regulators and lawmakers to understand relevant issues in the context of licensure and certification testing.

Once again, we are grateful to the authors and reviewers for contributing to our collective toolbox and enhancing our understanding of assessment in credentialing. Read, learn, and enjoy . . .
This column features a group of articles describing job analyses and papers on validity concepts. Topics covered in other articles include assessment using multiple measures, research on equating links, public test item disclosure, applications of computer-based testing to capture greater fidelity of work-related tasks, and prediction of test scores on the U.S. nurse licensure examination.

**Job Analysis**


This pair of articles contains a highly detailed description of methodology and findings from a role and function study conducted for the certified case manager (CCM) credential. Part 1 focuses on the job analysis study itself, including a survey that was sent to 52,370 individuals, yielding 7,668 usable responses. Part 2 follows the subsequent work on revision of the test specifications for the certification.

The Commission for Case Manager Certification (CCMC) conducts a role and function study every five years and uses the survey technique. The initial phases of the process are listed, starting with, “Using current literature and past similar study instruments about the role of the case manager, researchers draft an initial list of essential activities and knowledge areas, as well as a set of questions to ascertain the background and demographic characteristics of those participating in the study” (p. 271). Second, the draft description is “refined through input” from a representative group of practitioners. Third, a pilot study is conducted to obtain “input and feedback” on the survey. Finally the survey goes live nationally, emphasizing “important activities and knowledge areas” on which the certification examination should focus. The survey itself included 19 background and demographic questions, a list of 125 activities organized across six domains, and 94 knowledge statements organized across five domains. Both knowledge and activity statements were rated for importance and frequency (of activity performed or knowledge used) during an average day of work. A fourth section of the survey asked for ratings of adequacy of content (very poorly=0 to very well=4) and an opinion on the distribution of questions on the examination by domain. Finally, two additional opportunities were provided for open-ended comments.

Given the variety of positions associated with case management, demographics included 35 job titles and 30 work settings. Titles and work settings were combined based on similarities to obtain 16 groupings of job titles and 15 groups of work settings. A variety...
of analyses of demographics and activity and knowledge statements were performed.

A threshold of 2.5 was used for inclusion based on importance ratings, with 2.5 being the mid-point between “moderately important” (2) and “important” (3). “Overall, 106 items (84.8%) of the 125 essential activity items demonstrated 2.5 or greater mean importance ratings, and 68 (72.3%) of the 94 knowledge items demonstrated acceptable mean importance ratings” (p. 280).

The second (Part 2) article describes the process of development of revised content specifications for the examination. Eleven subject matter experts convened over a two-day period to determine the activity and knowledge statements that would be included in the content specifications and the weighting of the content domains. The average rating of 2.5 for importance ratings was a benchmark criterion, but the subject matter experts deliberated to come to consensus on the final selections, and those statements that were near the cut point were also considered. Ultimately 115 of 125 activity statements were included in the test specifications, and 81 of the 94 knowledge statements were included. An example of the deliberation included the decision that statements in the rehabilitation and vocational strategies domain should be part of the test specifications. Four of five statements were rated above “moderately important” (2) and a fifth statement had an average rating of 1.95.

To facilitate the analysis of data for the test specifications, analysis by respondent subgroups was performed, and a principal component factor analysis was used to focus attention on important elements of the data. As a result of the exploratory factor analysis, new domain names were developed for essential activities and knowledge based on the statements most highly associated with the factors. Finally, the subject matter expert panel members individually assigned a percentage weight for the knowledge domains. The panel results were compiled, and consensus was derived for the weighting of each knowledge domain. Knowledge domains were used as the basis for the table of specifications because “certification examinations test knowledge necessary for effective and competent performance in one’s role rather than the frequency and type of activities one engages in” (p. 18). The authors highlight two changes in the examination in two domains as a result of the role and function study. Quality management and ethical/legal practice were elevated to the domain level in the content outline.

This paper describes in detail the job task analysis performed by the National Consortium for Credentialing Health and Wellness Coaches (NCCHWC), a group that has been formed over a number of years to obtain data on the scope of practice of health and wellness coaching (HWC) and award a credential in this area based on best testing practice. In light of “heterogeneity in terms of role delineation and competencies,” a panel of 15 practitioners representing diversity of scope of practice and education was convened to devise a survey instrument. With the backgrounds of practitioners and the scope of practice known to be diverse, a literature review was conducted, including 284 articles. This review was based on the “explicit objective of clarifying the operational definition of HWC as it was emerging in the peer reviewed literature.”

To build the task list for the job analysis, a pool of 34 subject matter experts was nominated. From this group, 15 were selected to participate in the process based on obtaining the most diverse group possible. A 2-½ day meeting was held to develop a task list for the certification and achieve consensus on the tasks included. As described in the article, some limited discussion time also was devoted to knowledge and skills required of the HWC. “The Board convened a separate SME group composed of HWC educators with representatives from 20 different programs in academia, government, associations, and the private sector” (p. 3). This group developed the knowledge and skill statements to be used for training and education.

An online survey was developed that included demographic questions as well as importance and frequency ratings for the 17 listed tasks. The job analysis consultant/facilitator recommended adding four tasks for legal and professional considerations. Contacts for the survey invitations were established through organizations, training programs, and individuals. 885 responses were obtained for the survey. Importance ratings were based on a four-point scale (1=not important, 4=very important). A four-point scale was also used for frequency data (1=never, 2=infrequently—less than monthly, 3=occasionally—1-3 times per month, 4=frequently—greater than 3 times a month). A formula was used for combining frequency and importance data into a single weighted number for tasks. For the domain percentage distribution, the average percentages for panel ratings and the respondents’ ratings corresponded very closely.


This study by the National Board for Certification and Recertification for Nurse Anesthetists builds on previous
As a pilot for the survey, 65 nurse anesthetists were asked to provide review and comment. For the final version of the survey, a random sample of 12,000 nurse anesthetists were sent email invitations. Approximately half opened the email and 2,184 clicked the link to the survey (p. 44). 726 completed at least 50% of the survey. Means were calculated for both criticality and frequency. In addition responses were calibrated using Rasch measurement in order to develop interval scales. A subsequent meeting was held to finalize the content outline. The authors indicate that only minor changes were recommended. The changes included deleting the common topics domain and associated elements. Several methods were presented in the article to obtain domain weights combining criticality and frequency. The panelists ultimately approved domain weights based on the average of the three methods.


This is a relatively new medical specialty certification. For the years 2005 to 2013, examination content was defined by a committee of content experts. In 2014 a job analysis was performed. “The aim of the job analysis process was to produce examination specifications that accurately reflected the scope of endovascular practice, allowing for the development of a fair, accurate, and realistic assessment of each candidate’s readiness for certification” (p. 246). The kickoff activity for the job analysis process was a set of one-hour interviews with four designated subject matter experts (SMEs). “Interview questions solicited SMEs’ experiences and perspectives regarding regularly-performed job tasks, competencies required to perform job tasks, differences among practice settings, and the educational background of practitioners” (p. 246). Task and knowledge statements were then drafted. In a two-day meeting, a committee of eight SMEs refined the task and knowledge statements, organized them into a domain/subdomain structure, and made a preliminary judgment of weightings for content subdomains. The next step was a survey. Three scales were used for rating: frequency (“not relevant” to “very frequently”), importance (“not relevant” to “critically important”), and if required for entry level specialty practice (“not expected,” “some responsibility,” or “full responsibility”).

The population to be surveyed was defined as the group of individuals who were eligible to sit for the endovascular examination from 2008 to 2012. Of these 658 individuals, 126 (19.1%) responded to the survey. Domain weights were obtained from the job analysis survey, but adjustments were made for the examination blueprint based on a panel of SMEs’ review of the job analysis findings and “adjustments to reflect patterns amongst practicing endovascular physicians” (p. 247). In all, 110 test items were distributed among four domains, a number of subdomains, and specific diseases and disorders.

Validity Concepts

These days, when you think that you understand validity, you have to think again. A special issue of Assessment in Education: Principles, Policy & Practice is dedicated to validity concepts. One paper by Sireci (2016) has been previously cited in this column, but there are a number of others that are referenced below. The reader is reminded that the title of Sireci’s paper (“On the validity of useless tests”) communicates his position effectively. Of what use is a test for which there is no intended use?


The editors state that despite the appearance of agreement on the use of the term validity in published measurement standards, “...there is no widespread professional consensus concerning the best way to use the term” (p. 173). They state that the “controversy over consequences still looms largest” (p. 174). The closing question of the editorial offers a perspective that is optimistic: “Do we understand the issues at stake far better now than we did at the end of the twentieth century? We think that we do and believe that this special issue has made a significant contribution towards this end. The debate continues” (p. 176).


Consistent with the editorial that was co-authored by Newton, the premise of this paper is that there is no agreed-upon definition of the term validity in the literature. Newton
and Shaw offer three possible solutions to this issue, each of which would require the consensus of stakeholders regarding validity: “to eliminate its ambiguity by agreeing a precise technical definition, to embrace its ambiguity by agreeing a catchall lay usage, and to retire ‘validity’ from the testing lexicon” (p. 178). At the conclusion of the paper, embracing ambiguity of the use of the term validity and retiring the term from use are described as “leftfield” options. The authors suggest that they may sound “faintly ridiculous,” although “[t]hey invite us to consider a radically new discourse, either with or without the most hallowed term in our current lexicon. Yet, the semantic anarchism of the current situation is equally ridiculous. Radical times may require radical action” (p. 191).


Kane states, “I think of validity as the extent to which the proposed interpretations and uses of test scores are justified. The justification requires conceptual analysis of the coherence and completeness of the claims and empirical analysis of the inferences and assumptions inherent in the claims” (p. 198). “(L)icensure testing programmes are validated in terms of how well they assess knowledge, skills, and judgment (KSJs) critical for entry level practice. The goal is not to pick the best candidates, but to provide assurance that new practitioners have mastered the KSJs needed for effective practice” (p. 200).

Kane supports an argument-based approach to validation. “First, the interpretation and use is specified as an interpretation/use argument (or IUA), which specifies the network of inferences and assumptions leading from test performances to conclusions and decisions based on the test scores. Second, the IUA is critically evaluated by a validity argument” (p. 201).


Cizek refers to the common use of the term validity—interpretation of score meaning and support for the intended use of the scores—and declares the two types of evidence required for these functions to be incompatible. He states that an integrated evaluation framework is not possible because the two aspects of the term validity imply different questions to be researched. In fact, he refers to a fatal flaw in even the most recent standards documents resulting from the merging of interpretation of scores and gathering evidence for the appropriate use of the scores. In Cizek’s words, “The envisioned synthesis of evidence bearing on the accuracy of test score inferences and evidence bearing on the appropriateness of test score use is neither logically nor practically possible” (p. 214).

Cizek presents a table distinguishing between “sources of evidence for validating score meaning” and “sources of evidence for justifying test use.” For score meaning, there is evidence based on (1) test content, (2) response processes, (3) hypothesized relationships among variables, and (4) test development and administration procedures. In contrast, evidence for justifying test use is based on (1) consequences of testing, (2) costs of testing, (3) alternatives to testing, and (4) fundamental fairness (p. 220). Moss, P.A. (2016). Shifting the focus of validity for test use. Assessment in Education: Principles, Policy & Practice 23(2), 236-251.

Moss argues for “a more complex theory of validity that can shift focus as needed from the intended interpretations and uses of test scores that guide test developers to local capacity to support the actual interpretations, decisions, and actions that routinely serve local users’ purposes” (p. 236). She states, “By focusing validity theory exclusively on intended interpretations and uses of test scores, we overlook an important aspect of how tests actually inform and impact teaching and learning. Intended interpretations from standardised tests are always locally mediated and provide, at best, partial answers to local questions” (p. 247).


Markus states that the Newton and Shaw “prescriptive questions pose pseudo-problems, ill-defined questions having no constructive answers. A full presentation of my arguments for that conclusion proved too lengthy and technical for the special issue” (p. 252). The resulting paper is still quite technical; therefore, the excerpts in this abstract are acknowledged to be oversimplification. Markus uses “the unqualified term ‘validity’ to refer to test validity (i.e. validity in the context of testing) without proposing a specific object (e.g. tests). As a technical term, I distinguish test validity from validity of theorems in an axiomatic system, validity of logical arguments, validity of research designs and other common uses of the term” (p. 252). Four case studies are analyzed using the perspectives of different authors (e.g., Kane, Sireci) having what Markus describes as a definition of validity “inconsistent with the others” (p. 253).

Concluding comments in the paper are as follows: “Greater sensitivity to alternative vocabularies can help reduce talking past one another in the literature and facilitate constructive and effective criticism of existing theories. A focus on arguments that translate between vocabularies rather than efforts to impose or enforce a given vocabulary can more effectively promote progress in test validity theory….what each of us has to say about validity holds more interest that [sic] the preferred vocabulary each of us uses to express it” (p. 266).
Tests and Test Items

Although this is a measurement article, it is written in a reader-friendly style and covers important issues that affect decision makers in certification or licensure testing—observed score vs. underlying true score, error probability on pass-fail decisions, cost versus benefit of false-positive and false-negative decisions (based on discrepancy between observed score and the candidate’s underlying true ability), and the impact of using multiple measures for a certification decision. With respect to the latter topic, illustrations are provided in which a candidate is required to possess two different traits, each measured by a different assessment. As the candidate knowledge or attributes are different for the two areas, two different measures may be required. It is not always appropriate to use a model in which strength in one area can compensate for weakness in another. On the other hand, if there are two hurdles to be jumped, the odds of passing both are less than for either of the individual assessments. An example is cited in which a candidate has a probability of .70 of passing one assessment and .60 of passing a second assessment. The probability of passing both assessments is .60 × .70, or .42 (42%).

For their concluding remarks, the authors focus on recommendations to avoid false positive results—certifying a candidate who should not be certified. They state that they will “give it a try” to “sum up conclusions in a bumper-sticker aphorism” (p. 27). They make three suggestions with respect to an assessment consisting of multiple measures:

1. Restrict re-examinations to only taking the entire set of exams.
2. Scores are the average taken over all of the times the exam was taken.
3. Advise candidates that their own best interests might be best served by following the dictum, “If at first you don’t succeed, seriously consider an alternative profession.” (p. 27)

As a personal comment, the recommendations seem to be a bit harsh, but many of those in testing will be aware of programs where there is a single assessment with compensatory scoring for different sections, and a small number of candidates fail the assessment numerous times. Regardless of the low level of ability of the candidate, the odds eventually stack up in the candidate’s favor, not the public’s.

This summary by no means does the article justice. Time spent reading the article itself will be informative for individuals involved in certification and licensure testing.


For equating tests with anchor links of common items, the rule of thumb is that the anchor links or minitests (perhaps a quarter of the total test length) should be as much like the total test as possible; however, the full test frequently includes some items of extreme levels of difficulty or other aspects of less desirable statistics. This article brings together two topics: IRT (item response theory) equating and linking of test forms with miditests, anchor sets having the same level of difficulty as the total test but not the same distribution of item difficulty. A number of simulation studies were performed. The authors conclude, “We found that anchor tests with less variability in item difficulty than that of the full test performed about as well as or better than traditional minitests for most cases. Miditests appeared to be relatively robust to large differences in group ability, and may help to offset poorly estimated item parameters when the items on the test are not well targeted to the true ability of a group of examinees” (p. 188).


This article briefly reviews literature on item disclosure in medical licensure testing dating back to the 1970s in the U.S. but notes also that item disclosure issues are of current interest to countries such as South Korea, which maintains a fixed passing score at sixty percent correct. The authors look at (1) fairness and validity, (2) impact on passing levels, and (3) utility of item disclosure. In cases argued based on the public’s need to know, the authors indicate that 28 states have initiated legislation to require item disclosure, but only bills in California and New York have passed. Laws in these two states have, nonetheless, impacted both admissions and licensure testing.

The authors conclude, “While there are some noted advantages of test item disclosure—fulfilling the public’s need to know on test information, allowing test-takers to check the right answer and verify whether test results were processed properly, and enabling experts to examine items for potential bias—it may lead to more drawbacks and problems.” Limitations on testing programs include “making it impossible to conduct comparative analyses of each test based on test equating. Moreover, item disclosure hinders the reuse of test items; and as such, developing new items every year would require considerable financial and human resources.”
This paper is primarily focused on large scale computer-based assessments, the relationship of learning to assessment, and assessments created by organizations such as the Common Core State Assessment consortia, the Partnership for the Assessment of Readiness for College and Careers, and the Smarter Balanced Assessment Consortium. There is brief reference to innovations in certification and licensure testing. The author cites studies that “suggest that it is possible to build and administer assessments consisting of complex simulations and interactive performance tasks . . .” and adds, “Without question, the most highly developed use of these tasks for such decision making is found in the field of licensure and certification” (pp. 388-389). Three examples are cited: the American Institute of Certified Public Accountants (AICPA) Uniform CPA Examination, the National Council of Architectural Registration Boards (NCARB) Architect Registration Examination, and the U.S. Medical Licensure Examination (USMLE) administered by the National Board of Medical Examiners.

In addition to many multiple-choice questions, the Uniform CPA examination involves 23 performance tasks, described as “20 relatively constrained ‘task-based simulations’ . . . and 3 slightly longer written communication tasks” (p. 389). The NCARB Architect Registration Examination also has hundreds of multiple-choice questions and “11 complex performance tasks called ‘graphic vignettes’ which take from 1 to 4 hours each” (p. 389). In addition to over a thousand multiple-choice questions, the three steps of the USMLE contain “from 21 to 24 performance tasks. Between 9 and 12 of these performance tasks are computer-based simulations administered in a single 4-hour session . . . . The remaining performance tasks involve interacting with live simulated patients . . . ” (p. 389).

Bennett states, “Several common features of these three examination programs should be noted because they have implications for consequential educational testing. First, each program contributes to the making of a single proficiency decision . . . . Second, . . . multiple test parts taken over a significant period are aggregated to produce that proficiency classification. Third, complex simulations or other interactive performance tasks are included in each examination program. Those tasks are included because they are intended to measure competencies that a traditional test could not easily assess . . . “ (p. 389). Also mentioned are the resources for the candidate that go beyond a traditional test—tax codes or patient diagnostic information. The additional common features mentioned by Bennett are the number of performance tasks (at least 11), the fact that candidates must pass all sections, and the expenses of the tests, which are borne by the candidate. In all, the large number of traditional test items and performance tasks lead Bennett to conclude, “One obvious lesson is that to cover a domain broadly and deeply requires both a very long test and a combination of task formats” (p. 390).

**Prediction of Academic Success and High Stakes Test Performance**


This paper is a retrospective study on completion of a Bachelor of Science in Nursing (BSN) program and passing the nurse licensure examination (NCLEX) on the first attempt. The study is based on Seidman’s formula, which indicates that at-risk students can be successful in BSN programs if they can be identified early and given the necessary academic support. The study of predictor variables was based on two classes that were scheduled to graduate in 2010 and 2011, a total of 187 (predominantly young female Caucasian) students having complete data sets. The study consisted of a correlational study, with some additional analysis based on success rates projected by different levels of performance on admission variables and in-program measures. Predictor variables include pre-program grade point averages (GPA), ACT scores, anatomy and physiology course grades, and Health Education Systems Inc. (HESI) Exit Examination scores. The author concluded, “A significant relationship (p < 0.01) was identified between the preprogram GPA, ACT scores, anatomy grades, and the HESI Exit Exam scores with the completion of the BSN program and passing the NCLEX-RN.” Although no claim for generalizability of the findings to other BSN programs is made, the author is optimistic that “measures can be used to prevent nursing student attrition and promote completion of the nursing program and passing the NCLEX-RN for all students” (p. 223).

**Reference**

The Americans with Disabilities Act (ADA) and its subsequent amendments were designed to be comprehensive, civil rights type federal legislation providing opportunities for persons with disabilities. Signed into law in July 1990, the purpose of the ADA is stated as follows:

(1) to provide a clear and comprehensive national mandate for the elimination of discrimination against individuals with disabilities;
(2) to provide clear, strong, consistent, enforceable standards addressing discrimination against individuals with disabilities;
(3) to ensure that the Federal Government plays a central role in enforcing the standards established in this chapter on behalf of individuals with disabilities; and
(4) to invoke the sweep of congressional authority, including the power to enforce the fourteenth amendment and to regulate commerce, in order to address the major areas of discrimination faced day to day by people with disabilities. (SEC. 12101 [Section 2])

Under its various titles, the ADA prohibits discrimination on the basis of disability in employment, state and local government, public accommodations, commercial facilities, transportation, and telecommunications. It also applies to the United States Congress. The ADA applies to government activities and, under certain circumstances, to the private sector. In short, providers of programs and services must provide reasonable accommodations to otherwise qualified disabled persons to ensure equal access to such programs and services. States that issue licenses to qualified applicants under statutorily enacted criteria are required to provide reasonable accommodations to otherwise qualified applicants for licensure. Private sector entities that provide programs and services to the public must also comply with the ADA in the form of reasonable accommodations. Under some circumstances, the private sector assessment mechanism (such as an examination) is a mandatory criterion as a prerequisite to a state issued license.

Further complicating an already byzantine system, persons seeking accommodations under the ADA must substantiate the disability in order to determine eligibility and what accommodations will be provided. In supplying substantiating documentation of a disability, issues of privacy and confidentiality may surface. Finally, persons educated in a program outside the United States and not recognized or accredited by the relevant accrediting entity may have to substantiate educational equivalence with that of a United States education program. In some cases, foreign graduates may seek accommodations related to the establishment of education equivalence. Consider the following.

To become a licensed physician in any United States jurisdiction, applicants must meet eligibility criteria set forth in law. As one component of substantiating eligibility, and in addition to educational criteria, applicants must complete an examination developed to assess one's entry-level competence. For efficiencies and consistency purposes,
state boards have relied upon a uniform examination program as a means to assess entry-level competence. In human medicine, the examination relied upon by the state medical licensing boards is the United States Medical Licensing Examination (USMLE). The USMLE is developed, administered, scored, and maintained through a joint agreement between the National Board of Medical Examiners (NBME) and the Federation of State Medical Boards (FSMB). (Collectively, the NBME and FSMB may be referred to as Defendants.) Along with other requirements related to establishing education equivalence, graduates of foreign programs also must complete the USMLE as a prerequisite to licensure.

An applicant for access to the USMLE (Applicant or Plaintiff) filed litigation in the United States District Court for the Middle District of Pennsylvania against the NBME and the FSMB alleging discrimination claims under the ADA and the Rehabilitation Act. The Applicant is a 2004 graduate of a foreign education program who sought testing accommodations in 2006 and again in 2014. He failed the USMLE on six previous attempts and challenged the refusal of the NBME and FSMB to waive a six attempt limit on the examination. The Applicant, acting as his own lawyer, asserts state law contract claims and tort claims and seeks injunctive relief as well as damages. His compensatory and punitive damage claims are premised upon pecuniary harm and emotional distress suffered as a result of the alleged discriminatory conduct of the NBME and FSMB.

As can occur with litigation, the parties may ask the court to intervene on procedural issues related to discovery and the relevance of documents to the dispute. In this case, the Court was asked to address competing requests for the production of documents by the parties served upon each other.

First, the Defendants sought the release of the Applicant’s medical, psychiatric, education, and work records. To effectuate their discovery requests, the Defendants asked the Applicant to sign releases authorizing third parties to provide the requested medical documents. In the absence of a release, the Defendants indicated their intent to issue subpoenas to the relevant entities seeking the medical information.

On the heels of this exchange between the parties, the Applicant sent three proposed subpoenas to the Defendants seeking the personnel files of three current or former NBME employees. Shortly thereafter, the Applicant sent two additional subpoenas seeking the personnel files of two additional NBME employees.

The Applicant filed an objection with the Court alleging the Defendants’ subpoenas improperly disclosed his personal identifying information to the recipients and that the information sought was irrelevant. Based upon the pro se status, the Court generously interpreted the filing to be that of a protective order.

In January 2016, the Defendants filed a motion to compel compliance by the Applicant with the request for an authorization to release the medical information. The Defendants also sought a protective order regarding compliance with the Applicant’s requests for the personnel files of the five NBME employees. That same day, the Applicant filed his own motion for a protective order from compliance with the disclosure of his medical records. Finally, in March 2016, the Defendants advised the Applicant of their intent to serve several more subpoenas to third parties seeking additional medical, psychiatric, and education records.

The Court was asked to rule on the counter motions to compel compliance or, alternatively, issue protective orders to prohibit enforcement of the subpoenas. The Court identified four issues:

1. Whether the Defendants’ subpoenas improperly disclose personal identifiable information;
2. Whether the Defendants’ subpoenas are unenforceable as requesting irrelevant information;
3. Whether the Applicant should be compelled to sign authorizations to release his medical, psychiatric, and education records; and
4. Whether the Applicant’s subpoenas are unenforceable as requesting irrelevant information.

Regarding the issue of disclosure of personal identifiable information, the Court noted the importance of ensuring that the information requested and provided involves the appropriate patient. Generally, discovery materials are not filed in the public court record but only in the private files of the parties and counsel. To the extent any future use of the information is necessary and relevant to the judicial proceedings, the federal and local rules provide for redaction of sensitive information. In addition, attorneys can seek the filing of materials under seal. Thus, the Court found the Applicant’s arguments to be without merit and overruled his motion.

Addressing the relevance of the medical, psychiatric, and education records of the Applicant, the Court held that the Applicant filed a lawsuit under the ADA claiming that
his mental impairment, bipolar disorder, and previous misdiagnosis of attention deficit/hyperactivity disorder places him in a protected category subject to the relevant federal laws. He seeks damages based on mental anguish, subversion of his self-esteem and self-worth, and claims that the actions of the Defendants affected his overall health. He also claimed that such actions caused him to have a nervous breakdown, exacerbated an existing condition, and imposed significant stress on his well-being. Allegations in support of his damages also included headaches, high blood sugar, skin conditions, anxiety, and sleep disturbance.

Under applicable rules of discovery, the Defendants are entitled to access information that is relevant to any claim or defense and “proportionate to the needs of the case.” The Court found that any patient-mental health practitioner-physician privilege that might have existed had been waived by the Applicant placing his physical and mental health at issue in the case. Plaintiffs cannot be permitted to hide behind a claim of privilege when such condition is central to the dispute. Consequently, the Court held that the Applicant’s medical and psychiatric records are relevant and subject to discovery. Also, based upon the claims of lost income and diminished employment opportunities, the Applicant’s education and employment records are relevant and subject to discovery. Finally, the Court rejected the notion that the potential for embarrassment justified the prohibition from disclosure. Consequently, the Court overruled the Applicant’s motion and held that the requested information is subject to disclosure.

As a means for the Defendants to gain access to the requested medical, psychiatric, and education records, they sought an order from the Court compelling the Applicant to sign releases to the relevant record holders. Based upon the above analysis, the Court held in favor of the Defendants and found that the Applicant is compelled to promptly sign the authorization forms.

Turning its attention to the discovery requests of the Applicant, the Court addressed the requests for the personnel files of the five current or former employees of the Defendants. The Applicant argued that the Defendants failed to accommodate his alleged disabilities in administering the USMLE. Specifically, the Applicant asked for and was denied extra time in 2006 and sought a waiver of the six attempts rule after his failed attempts at the examination. Agreeing with the Defendants, the Court found that the personnel records are not relevant to the decisions to deny the requested accommodations. The Applicant had not articulated any reason why the confidential personnel files are or could be relevant to the institutional decisions of the Defendants. Accordingly, the Defendants’ objection to the production of these files was sustained, and the files were deemed not subject to disclosure.

The Court allowed the Defendants access to the requested medical, psychiatric, and education records while denying the Applicant access to the personnel files of the five employees.

As examination entities are assessing exam candidates under the ADA, an analysis and substantiation of the disabilities is necessary to negotiate the accommodations required or desired to be provided. Failure to provide requested accommodations may lead to litigation by the dissatisfied candidate. At both the assessment and litigation stages, the relevance of the medical records is likely at issue. Submissions and, where necessary, waivers from the candidates may be appropriate to make informed decisions regarding the exam administration. Security and validity of the exam results as an indicator of entry-level competence are of critical importance to the exam entity when navigating these decisions.

References


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 annual CLEAR 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 that 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 cer@clearhq.org. Please enter “CER Perspectives on Testing” in the subject line.

Should Testing Organizations Report Sub-scores?

Response provided by Peter B. Mackey, CFA
head, exam development
CFA Institute

Score reports are an important component of quality licensing and credentialing programs and a key ingredient in customer service. Credentialing organizations are always looking for ways to improve reporting exam results. Where exams cover different subject matter, topics, or domains, sub-score reporting (by domain level) is an area worthy of consideration. Score reports can assist unsuccessful candidates in identifying and understanding weaknesses, considering retesting, and understanding where to focus future study efforts. Motivations may also include providing guidance for continuing education or exam preparation in related knowledge/skill areas.

Like most testing concepts, it’s never simple. Experts raise a number of caution flags for sub-score reporting. The Standards for Educational and Psychological Testing (2014) emphasize that any information provided should be explained in a manner such that the
intended interpretation, usage, and reliability are clear to the users (Standard 6.10). Any recommendations regarding improving future performance should be based only on empirical evidence. The National Commission for Certifying Agencies (NCCA) Standards for the Accreditation of Certification Programs echoes this guidance: Report only reliable performance indicators and disclose any limitations in interpreting the data (Standard 20).

Candidates want as much performance information as they can get; testing organizations are responsible, however, for ensuring that users of the information can rely on it to achieve the stated purpose. If psychometric evaluation indicates all exam topics are related to a single construct (the concept or characteristic that a test is designed to measure), it may be best not to report sub-scores. It would be better to report only overall performance than weak or potentially misleading indicators. Candidates are entitled to assume that a sub-score is an accurate indicator of their knowledge, skills, and abilities in that domain.

Exam length, the number of topic domains, and the number of exam items in each domain are all useful considerations in deciding how much information to provide. My view is that if the domain is large enough to support reliable performance indicators, it is best to provide the sub-score. Even if a psychometric study concludes the domains are subsumed in a larger single construct, practitioners may consider the domains separate areas of knowledge.

The Chartered Financial Analyst (CFA) exams developed by CFA Institute are organized into three progressive levels; each level covers 10 domains in a 6-hour exam. Although the domains are shown to belong to a single construct, we have chosen to report candidate performance in each domain. Investment professionals and CFA candidates consider the subject matter in Financial Reporting and Analysis, for example, to be very different than the material covered in Fixed Income Investment Analysis. Looking forward, we are considering merging some closely-related topics for performance reporting purposes. Our intention is to improve the reliability of information we provide to candidates, particularly with regard to the lower-weighted domains.

If, alternatively, the exam is relatively short or topic domains make up a small proportion of the test, sub-score reporting can communicate misleading signals. Even where the testing organization is confident in the reliability of sub-scores, it may be advisable to tell examinees that sub-scores should be considered estimates of ability level, not necessarily reliable predictors of future performance in the respective domains.

In summary, I recommend that you consider reporting sub-scores on your exams for any domain that represents a significant portion of the targeted knowledge and skills of the profession. Importantly, you will need to enlist the support of your psychometric staff or consultants to ensure the data you are reporting are reliable and sound. Lower-weighted domains can often be combined as long as the content is closely related. Candidates, especially those whose exam efforts have been unsuccessful thus far, will greatly appreciate sub-score information and use it to support future development.

**References**


**Should we allow “easy” items on the exam?**

*Response provided by Ellen Julian
senior director psychometric services
Inteleos*

Test items may be considered as too easy, too hard, or of appropriate difficulty for measuring candidate knowledge. While hard items may indicate content problems, easy items may represent, at the worst, a missed opportunity to gather additional relevant information about how much the candidate knows.

For multiple-choice questions (MCQs), we tend to define “too hard” as anything close to, or below, the proportion-correct (p-value) expected from people selecting responses at random (i.e., 0.25 for four-choice MCQs). However hard, MCQs’ p-values seldom approach 0.00. Hard items with low or negative discrimination (typically point-biserial) values may be eliminated automatically, even if no content-related reason is obvious, just in case the item is flawed (e.g., ambiguous, more than one correct answer).

Easy items, in contrast, are almost certainly keyed correctly and the candidates agree on which response option is the correct answer. Some consider an item easy if more than 95% of the candidates answer the item correctly. Ironically, easy items may have low discriminations because they are closer to the end of the p-value scale. If the handful of candidates who miss an easy item are not the lowest...
scorers on the entire test, the item’s discrimination may be so low that it is flagged as a problem. Don’t be too quick to condemn an easy item because of a low point-biserial.

“Too easy” may look different when deciding which items to score versus which items to select for a new test form.

When building new test forms with items of known difficulty, select items with difficulties near the cut score. If the items have been calibrated with an Item Response Theory model, this is straightforward because their difficulty is on the same scale as the cut score. With Classical Test Theory (p-values and point-biserials), a simple rule of thumb can be helpful: If a test passes X% of examinees, its most informative items will pass about that percent. For instance, a test that passes 80% of candidates should have an average p-value of about .80. Tests that are easy to pass should have easy items and tests that are hard to pass should have hard ones, so that performance on an item informs the pass/fail decision rather than making distinctions between people far from the critical pass/fail zone.

Content coverage trumps difficulty, however, so if the only available item in a required content area is easy, there’s no harm in using it. However, items of a more moderate difficulty (which often are better targeted to the cut score) would be preferable because they are adding to the measurement precision as well as meeting content requirements.

When a test form has already been administered and items are being selected for scoring, easy items should probably be scored, since the examinees’ time and mental energy has already been spent on them. While they may not be adding to the test’s measurement information, they’re not doing any harm. As always, content comes first, so if you need the easy item to meet content specifications, score it with no apologies. However, if you need fewer in its area, too easy items should be in line for deletion, right behind negatively-discriminating and too hard ones.

**Is there “bang for the buck” when implementing innovative items?**

*Response provided by Anthony Zara*  
*vice president, assessment solutions*  
*Pearson VUE*

This question is currently on the minds of stakeholders from many computer-based testing programs. It is a natural next step after transitioning an examination program from paper-and-pencil delivery to computer. Often, credentialing testing programs decide to transition to computerized delivery expressly to enable the use of innovative items. For the purposes of this discussion, “innovative items” means those examination items that utilize unique stimuli or processes—in the question presentation/stem, in the manner of responding, or in the navigation within an item. Although not particularly innovative, we would consider question types using a short video or animation in the question presentation or having a free response answer as being innovative.

Psychometricians often respond to questions by saying, “It depends.” The value of innovative items to a program depends on many factors that are unique to the program and on the motivation to implement these types of items. It is not at all a simple answer to determine whether or not using these types of items is generally “worth it.” However, there is a set of interesting questions that will provide some direction regarding the level of “worth it” when using or contemplating using these types of items.

The first question to ask is why are you implementing these item types? Potential reasons include:

1. These items uniquely tap an important part of the testing construct that multiple choice items (MCQs) cannot.
2. These items provide additional measurement information along with MCQs that will improve the validity of our pass/fail decisions.
3. My organization needs to present a “next generation” assessment to candidates to keep our stakeholders satisfied.
4. My board believes we need to modernize our examination to stay relevant.
5. Educators for my profession think that some important functions related to job performance are not adequately captured by MCQs alone.
6. The innovative items more closely mirror the types of questions asked during the educational process.
7. It will be really ground-breaking to implement these items.
8. Innovative items more adequately mirror aspects of the job than do MCQs.

Other psychometric questions that must be answered in order to determine the relative value of innovative items to your program include:

1. Are we able to produce innovative items in sufficient numbers to provide reliable and valid measurement on our examination?
2. Are we achieving an increase in important information about our candidates that will impact pass/fail decisions?

3. How much do the innovative items correlate with the existing construct as measured by our MCQs?

4. Can we integrate the innovative item scores with the MCQ scores to produce a valid pass/fail result, or will we need to develop a result with two scored components?

5. Do the innovative items show any adverse impact to identified groups of candidates?

There are also key operational questions to be answered in order to determine the relative value of including innovative items:

1. How much does it cost to produce the innovative items from a cost, staff time, mindshare, and distraction standpoint?

2. Is the organization prepared to mount a significant education campaign (for candidates, stakeholders, the public, etc.) before implementing the innovative item types?

3. Is the organization capable and willing to create a comprehensive candidate practice opportunity for these item types?

4. Is the organization set up to handle the increase in stakeholder queries about the exam and exam process?

As suggested based on the above, determining whether your program gets “bang for the buck” for implementing innovative items is a fairly detailed exercise. Several licensure programs have taken the leap to using innovative items; it might be informative to develop a dialog with their staff and stakeholders to learn about the possible pitfalls and sets of benefits that your program might expect.
Item Writing: Methods for Generating a Greater Return on Investment

BELINDA BRUNNER, M.S.
director, testing services strategy
Pearson VUE

Abstract

Item writing is an expensive exercise for most licensure organizations, and a significant portion of item development costs are associated with holding item writing and review meetings with subject matter experts (SMEs). While automated item generation (AIG) is an exciting development in high-stakes testing, the majority of examination programs still use some form of the traditional SME committee approach to item writing. This article provides a review and discussion of innovative techniques that can be introduced in the traditional non-automated item writing process to increase productivity and to use SMEs more effectively. It advocates for more research on effective item development strategies.

Keywords: item writing, item models, item writer productivity, item quality

Item Writing: Methods for Generating a Greater Return on Investment

Test content is an important – some would say the most important – asset for test owners. The cost of developing an operational test item for a high-stakes test has been estimated at $1500 to $2000 (Rudner, 2010). Item development methods used by high-stakes testing programs, such as licensure programs, involve the use of Subject Matter Experts (SMEs) to write and review items. While SMEs may write items remotely, there are almost always opportunities for face-to-face meetings for training or reviewing items if not for the actual writing. Costs associated with supporting SME meetings can be a significant portion of the expenditures associated with item development.

Automated item generation (AIG) has been offered as an alternative means of creating large numbers of test items. In AIG, item models or prototypical representations of items are the basis for automatically generating test items (LaDuca, Staples, Templeton, & Holzman, 1986; Bejar, 2002; Gierl & Haladyna, 2013). Although SMEs are still an integral part of the AIG process, their role changes and is no longer focused on authoring test items one at a time. SMEs define variables for elements in the item model that can be manipulated to create new items, and then computer technology is used to systematically combine the variables to make new items (Gierl, Lai, & Turner, 2012). In its most ambitious form, AIG would be used to produce statistically calibrated items from each item model (Gierl & Lai, 2013).

Computer-based testing (CBT) has become a popular test administration model for licensure programs. On-demand computer-based testing requires a greater number of test items to prevent over-exposure of test items. Complex test administration structures enabled by computer-based testing, such as adaptive testing, also have a greater item demand. While AIG offers a means of meeting the item demands required by adaptive testing and other complex test administration structures (Lai, Alves, & Gierl, 2009), it doesn’t yet enjoy widespread adoption by high-stakes testing programs.
The expense devoted to item development in exam programs and the increasing item demand requirements have made test owners look for ways in which productivity in the item writing process can be increased. Yet there is a scarcity of literature discussing innovative approaches to traditional item development techniques as well as little research available on the effectiveness of different techniques. This article looks at innovations within the traditional non-automated item writing process to increase productivity and to use SMEs more effectively.

Sourcing Items

The Traditional Committee Approach

Traditional item writing methods used by licensure programs frequently involve SMEs serving as a standing committee of item writers. The standing committee typically meets periodically to write and sometimes review items, and training is usually provided during these in-person meetings. Technology has provided newer enhancements to the traditional standing committee process. Web conferencing means that virtual meetings can be conducted in lieu of or as a supplement to face-to-face meetings. Remote authoring capabilities in item banking and content management software facilitate item writing outside of committee meetings.

Alternative Organizational Structures for Sourcing Items

Case, Holtzman, and Ripkey (2001) compared three approaches to item development for the National Board of Medical Examiners (NBME). One of these approaches, item harvesting, resembles crowdsourcing, which will be discussed in the next section, although Case et al. (2001) reported on item harvesting before the advent of the term crowdsourcing. Item harvesting was performed by asking individual physicians to write items and sending them instructional materials including an item writing manual. They compared this approach to the traditional standing test committee and an ad-hoc item writing task force. They found that more items were approved for live use per author and that items had a higher quality rating with the traditional committee approach than with the other two methods. Case et al. (2001) concluded that the ad-hoc task force method was also a top-quality item development model. They suggested that item harvesting might be a good choice for a low-budget program or to supplement the item writing process in a high-budget program.

Crowdsourcing

Crowdsourcing is the process of obtaining ideas, products, services, or content by soliciting contributions from a wider group online. Crowdsourcing is a relatively new concept, the term having first been used by Jeff Howe, a contributing editor to Wired magazine, in 2006. Crowdsourcing has been reported in the literature as a means of gathering content for professional issues, such as developing a knowledge database for linking clinical problems with medications (McCoy, Wright, Laxmisan, Ottosen, McCoy, Butten, & Sittig, 2013), conducting content analysis for managerial research (Conley & Tosti-Kharas, 2014), and designing mathematics curricula (Scott, 2015).

Within the context of test development, Sadler, Sonnert, Coyle, & Miller (2016) used a crowdsourcing approach for pre-testing test items for a middle-school life science assessment as a means of gauging item quality. Item performance data obtained from the sample of crowdsourced adults were compared to those obtained from middle schoolers. The authors found that correlations between the item parameters obtained from the adult sample and those obtained from the middle school sample were good. They concluded that crowdsourced data could be used to reject items with low item discrimination.

While Sadler et al. (2016) examined crowdsourcing as a means of evaluating item quality, could it also potentially be used as a means of sourcing items? Further study into the use of crowdsourcing within the item development process may be warranted. One potential use is in the identification of ideas upon which items can be developed. The critical incident technique is a commonly-accepted means of gathering data within a job task analysis to determine examination content domains. Critical incidents, common work situations, or common professional errors could be crowdsourced from professionals working in the field to serve as ideas from which items could be developed by trained item writers. Using crowdsourcing in this manner may lead to a greater range of concepts upon which items are developed while using the skills of trained item writers to finesse the ideas so that item quality is not jeopardized.

Tools and Techniques to Assist in the Item Development Process

Enhanced Instructions and Materials

Item writers typically receive item writer training, assignments on content areas for which items need to be written, and instructions on the particulars of writing items for the exam program, such as style guidelines. Becker and Olsen (2012) explored ways of providing item writers with more information regarding ideas upon which items could be developed. They used natural language processing and latent semantic analysis of electronic textbook materials to provide supplemental
information to item writers in the form of prepared materials per blueprint that included excerpts from textbooks to provide key information and glossary terms that were missing from the item banks. Item templates were also provided to item writers and discussed during a training session.

Item writers were divided into an experimental and a control group. The experimental group was given the supplemental textbook information and the glossary terms to assist in the item writing process, while the control group had the item templates but not the additional materials. Findings revealed that the experimental group wrote 23% more items than did the control group (Becker & Olsen, 2012).

**Item Models**

Traditional item development methods are costly in part because they treat items as “unique craft projects” (Luecht, 2013, p. 71) created one at a time through a piecemeal process. One of the key concepts of automated item generation is the use of item models (LaDuca et al., 1986), also called item shells (Haladyna & Shindoll, 1989) or templates (Mislevy & Riconscente, 2006). An item model is a type of item master or pattern with variable elements that can be changed to create new items. An example of an item model for an accounting item is presented in Figure 1. The variable elements are identified within the item's stem, and rules or constraints for generating alternatives for each element have also been defined.

Item models such as the one represented in Figure 1 can be used with an automated item generation engine (see Gierl & Haladyna, 2013) to generate item variants electronically through the use of computer software. However, in the absence of access to such software, the concept of item modelling can be used as part of an item development process solely using human capital. Figure 2 provides another example of an item model, this time for a logical reasoning item. Examples of three additional items created using the original item as the model are also presented in Figure 2. It should be noted that, unlike the model in Figure 1, editing is required for each resulting item stem to create items that are grammatically correct. Such editing can be performed by editors or proofreaders rather than the SMEs themselves, allowing the SMEs to devote their time to the components of the item development process that do require their subject matter expertise.

If the desire for using the item models is to create items that have comparable psychometric characteristics, then surface elements, or *incidentals*, should be manipulated to create new items. If the desire is to create items with different psychometric properties, then the deep characteristics, or *radicals*, should be manipulated (Gierl & Lai, 2013). One of the reasons for identifying the constraints to be placed on the replacement of elements, as illustrated in Figure 1, is to try to control the effect the manipulation of elements may have on the items’ psychometric properties.

Shea, Poniatowski, Day, Langdon, LaDuca, & Norcini (1992) reported on an adaptation to item modelling that was used with novice item writers. An item model was created from a source item by identifying the critical elements in the item. The adaptation was that then the item writers were asked to alter a single-key multiple-choice item so that 1) an incorrect answer option became the key, and 2) the question in the stem was changed. Although the study was limited to two item writers, the authors reported that the results suggested that the method used could be an efficient item-writing procedure and that it was about four times as productive as the traditional committee approach.

Regardless of whether an item model like the ones presented in Figure 1 or Figure 2 or as adapted by Shea et al. (1992) is introduced within an existing item development process, the goal of their use remains the same. One SME idea is used to generate a number of items. Unlike the "unique craft project" approach suggested by Luecht (2013) as common in traditional item development, there is a one-to-many relationship between idea and item rather than a one-to-one relationship.
Five ships \([\text{Element 1}]\) arrived in port \([\text{Element 2}]\) on five consecutive days \([\text{Element 3}]\), Monday through Friday \([\text{Element 3a-e}]\). The five ships are the Independence \([\text{Element 1a}]\), the Endurance \([\text{Element 1b}]\), the Catalyst \([\text{Element 1c}]\), the Prosperity \([\text{Element 1d}]\), and the Stalwart \([\text{Element 1e}]\).

The Prosperity \([\text{Element 1d}]\) was the last to come into port \([\text{Element 2}]\).

The Stalwart \([\text{Element 1e}]\) arrived in port \([\text{Element 2}]\) two days \([\text{Element 3}]\) before the Catalyst \([\text{Element 1c}]\).

The Endurance \([\text{Element 1b}]\) arrived in port \([\text{Element 2}]\) on Tuesday \([\text{Element 3d}]\).

What day \([\text{Element 3}]\) did the Independence \([\text{Element 1a}]\) arrive?

<table>
<thead>
<tr>
<th>Element Definition</th>
<th>Object of Item</th>
<th>Named object 1</th>
<th>Named object 2</th>
<th>Named object 3</th>
<th>Named object 4</th>
<th>Named object 5</th>
<th>Collection</th>
<th>Time frame</th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
<th>Time 4</th>
<th>Time 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>Ships</td>
<td>Independence</td>
<td>Endurance</td>
<td>Catalyst</td>
<td>Prosperity</td>
<td>Stalwart</td>
<td>Port</td>
<td>Days</td>
<td>Monday</td>
<td>Tuesday</td>
<td>Wednesday</td>
<td>Thursday</td>
<td>Friday</td>
</tr>
<tr>
<td>Variant 1</td>
<td>Planes</td>
<td>Chicago</td>
<td>Philadelphia</td>
<td>New York</td>
<td>Los Angeles</td>
<td>Houston</td>
<td>Airport</td>
<td>15 min</td>
<td>10:00 am</td>
<td>10:15 am</td>
<td>10:30 am</td>
<td>10:45 am</td>
<td>11:00 am</td>
</tr>
<tr>
<td>Variant 2</td>
<td>Classes</td>
<td>Art</td>
<td>Science</td>
<td>Geography</td>
<td>History</td>
<td>Music</td>
<td>Time slots</td>
<td>Sessions</td>
<td>1st session</td>
<td>2nd session</td>
<td>3rd session</td>
<td>4th session</td>
<td>5th session</td>
</tr>
<tr>
<td>Variant 3</td>
<td>Books</td>
<td>To Kill a Mockingbird</td>
<td>The Old Man and the Sea</td>
<td>Ulysses</td>
<td>Pride and Prejudice</td>
<td>Doctor Zhivago</td>
<td>Book club choices</td>
<td>Club meetings</td>
<td>January</td>
<td>February</td>
<td>March</td>
<td>April</td>
<td>May</td>
</tr>
</tbody>
</table>

**Item Variant 1:**
Planes from five destinations arrived at an airport in 15-minute intervals. The planes were travelling from Chicago, Philadelphia, New York, Los Angeles, and Houston. The first plane arrived at 10:00 in the morning.

The plane from Los Angeles was the last to arrive.
The Houston plane arrived at the airport 30 minutes before the one from New York.
The plane from Philadelphia arrived at 10:15.

What time did the plane from Chicago arrive at the airport?

**Item Variant 2:**
Classes held in consecutive time slots at a particular school were Art, Science, Geography, History, and Music. They were not held in that order.

History was the last class of the day.
Music was held two sessions before Geography.
Science was the second class of the day.

In what class session was Art held?

**Item Variant 3:**
A book club holds monthly club meetings. A different book is discussed each month. For the first five months of a particular year, the books discussed were To Kill a Mockingbird, The Old Man and the Sea, Ulysses, Pride and Prejudice, and Doctor Zhivago, but not in that order.

Pride and Prejudice was the last of the five books discussed.
Doctor Zhivago was discussed two club meetings before Ulysses.
The Old Man and the Sea was discussed at the February meeting.

In what month was To Kill a Mockingbird discussed?
**Idea Generation Techniques**

The genesis of every test item is an idea. Ebel (1951, p.190) asserted, “The production and selection of ideas upon which test items may be based is one of the most difficult problems facing the item writer.” Licensure programs typically have some content areas for which it is more difficult to develop items than others. Item writers can incur writer’s block in which they find it difficult to come up with new ideas upon which to base items.

Since there is a scarcity of literature that discusses innovative item development techniques, literature from other domains, such as creative thinking, can provide approaches that may be adapted to the item writing process. Two creative thinking techniques that have the potential to aid in the item writing process are the comedy workshop format and the SCAMPER technique.

**The comedy workshop format.** The comedy workshop is a format used to develop scripts for television in which writers work collaboratively to brainstorm ideas. Droegemueller, Gant, Brekken, & Webb (2005) reported on an adaptation of the comedy workshop format used within the item writing process for the American Board of Obstetrics and Gynecology. Item writers worked within groups to generate ideas and develop items. Amongst the advantages reported were that the workshop format aided in teaching inexperienced item writers, was an enjoyable experience for the item writers, and improved the quality of the items authored. However, the authors recognized that it was labor intensive and that it could be difficult to keep the item writing groups organized and on track.

**The SCAMPER technique.** The SCAMPER technique was first described by Eberle (1996) as a means of aiding creativity.
and innovation during the brainstorming process. SCAMPER is a mnemonic that stands for (S)ubstitute, (C)ombine, (A)dapt, (M)odify or (M)agnify, (P)ut to another use, (E)liminate, and (R)everse. Each letter in the mnemonic represents a different way of thinking about how a product or an idea could be changed or improved to generate a new idea or solution. Brunner and Rimington (2015) described how this process might be used in the item development process to brainstorm on ideas for new items using existing items as the stimuli. Figure 3 provides an example of the types of questions that might be asked using the SCAMPER technique with a pharmacy item to generate new items.

The SCAMPER technique can be used when an injection of new ideas is needed, such as for content domains for which it is difficult to write items or when item writers are exhibiting writer’s block. It can also be used with item models as a way of generating alternates for the elements of an item.

**Summary**

Automated item generation offers the promise of a more productive and efficient item development process. However, at the moment, its adoption may not be realistic for a large number of licensure programs, particularly smaller programs with lower budgets. This article has provided a review of the literature with regard to tools and techniques that can be used within the traditional SME committee approach to item writing to improve productivity and efficiency.

The goals to be achieved with the adoption of the tools and techniques presented include using SME time more wisely and generating more items in a shorter period of time. The experiences and preferences of item writers should be considered in deciding which tools or techniques are appropriate for a particular examination program. For instance, inexperienced item writers might find item models or templates helpful, while more experienced item writers might find that they are unnecessary or even cumbersome.

While productivity is an important goal, generating a larger number of items will only really make a worthwhile contribution to an examination program if there are a larger number of usable items. As Ebel stated, “Skill in item writing depends not only upon prolific inventiveness but also upon discriminating judgment in the selection” (1951, pp. 190-191). Some of the studies discussed have suggested that item quality may also be improved using the approaches presented. However, more research is needed on effective means for improving item writer productivity and item quality. In general, while there is a significant body of research on the psychometric aspects and procedures of test development, research on item development strategies is lacking. Such research would contribute to establishing an empirical foundation for determining effective item writing strategies.

**References**


Clarifying the blurred lines between credentialing and education exams

CHAD W. BUCKENDAHL, PH.D.
partner, ACS Ventures, LLC

Introduction

Psychological measurement is inclusive of a range of applications from cognitive to psychomotor to attitude. Generally represented as tests, tasks, and instruments, it is easy to recognize these measurement activities across multiple industries. Participants in these processes are just as diverse. Large segments of society encounter these data collection, analysis, reporting, and interpretation efforts from an early age, often beginning with participation in educational systems. Throughout the educational process, psychological measurement occurs at multiple checkpoints along this continuum. An intersection of measurement activities between educational and credentialing (e.g., licensure, certification) purposes starts to occur at the secondary level when students may have opportunities to encounter industry certification as part of post-secondary career pathways. However, the different purposes of these tests cannot be dismissed. These distinctions extend to the credentialing programs that may be more familiar to members of the public.

At the heart of any psychological measurement activity is the concept of validity, which is characterized by the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; hereafter Test Standards) as the collection of evidence that supports defined interpretation(s) and use(s) of scores. The Test Standards serve as expectations for professional practice and are inclusive of the range of psychological measurement activities noted above with specific discussions about the sources of evidence for different applications (e.g., clinical, education, credentialing, employment). Although there are core expectations within the concepts of validity, reliability, and fairness, there are some notable differences among these different applications and how they are implemented in practice that are worth discussion.

As lines blur among uses of test scores for education, credentialing, and employment, it is important for regulatory bodies, policymakers, and practitioners to understand some of the key differences among these uses and the types of evidence that might be prioritized to support a validity argument (Kane, 2006). The purpose of this article is to discuss three key distinctions among tests that are developed for educational as opposed to credentialing purposes. Although these differences should be interpreted as illustrative and not exhaustive, they represent some of the more frequently discussed questions by credentialing programs when faced with questions by stakeholders (e.g., candidates, education programs, policymakers).

The first substantive difference between an educational testing program and a credentialing program is the evidence of alignment of curriculum, instruction, and assessment (Webb, 2006; Davis-Becker & Buckendahl, 2013). “Curriculum” is terminology used

---

1 Within the article test, examination, and assessment are used interchangeably to represent a measurement instrument.
in an education context and can often represent potentially subtle distinctions of how content is defined and derived from learning objectives or academic content standards, as opposed to the credentialing context where this information is collected and evaluated through a practice or job analysis.

A second notable difference is that for credentialing programs, the focus of the measurement is most efficiently designed at the point of distinction between minimally competent or not. Conversely, many of the large-scale education assessments are designed to measure and classify students into three or more levels of performance. For credentialing, this emphasis around a single cut point does not necessitate the level of breadth and depth of coverage as in education.

The third difference discussed in this article highlights the concept of construct-relevant accommodations as a point of fairness and its contribution to validity. As an important disclaimer, the interpretation of accommodations will intersect with national or state-level jurisdictional laws or regulations. The author is not qualified to offer a legal opinion and instead would refer interested readers to Jacobs (2016) and their own counsel. However, the Test Standards provide valuable guidance that practitioners and policymakers can use when designing their programs and considering potential accommodations, as opposed to a modification, that candidates could request. Because the purpose of an accommodation is to facilitate demonstration of learning or abilities that do not interfere with interpretation of the construct, this distinction between an accommodation and modification should be explored.

In the sections that follow, each of these distinctions is discussed in greater depth to add clarity to selected differences between educational and credentialing applications of psychological measurement.

Alignment of curriculum, instruction, and assessment

“Alignment” can be defined as the qualitative and quantitative relationships among curriculum, learning objectives, instructional delivery, and assessment (see Webb, 2006, Porter, 2002). For education programs, evidence of alignment is often characterized as the extent to which there is overlap of content, cognitive processing, and performance demand as intended in these components of the educational system (Davis-Becker & Buckendahl, 2013).

The instructional delivery component of this system in the educational context further highlights an important difference from a credentialing context. Specifically, the delivery of instruction is a source of evidence characterized as opportunity to learn. This means that beyond just the evidence that the depth and breadth of the content is included in the curriculum, it is evidence that the content was delivered through instructional activities (e.g., lesson plans, observed instruction, classroom assessment) that support learning.

This concept was highlighted in Debra P. v. Turlington (1979, 1981, 1983, 1984), a case involving graduation examinations in Florida and differential instructional policies and practices that previously occurred in the state. Although graduation examinations have characteristics of a credentialing examination (i.e., diploma eligibility), the intended inferences from scores on these examinations reflect the educational process and outcomes and not any given job role or occupation.

Although alignment as a concept also applies to credentialing programs, the connections between the curriculum and instruction when compared to the test are not critical, particularly when there are multiple pathways that may lead to eligibility or preparation for a credentialing examination. Rather, the important source of alignment evidence for credentialing programs will be the extent to which the evidence from the practice analysis corresponds to job-related knowledge, skills, and abilities.

In the educational context, the evidence for education assessments is less determined by what is observed in classroom practice, but rather by what is desired as learning outcomes for students.

Focus of measurement

The second area of discussion is the focus of measurement and how it aligns with the purpose of the examination and the scores that are reported. Philosophies in education about how to best assess and report student achievement have ebbed and flowed as interpretation and uses have changed.
Practices often seen in classroom assessment are ones that apply student performance to an arbitrary scale (e.g., 70% = C, 80% = B, 90% = A) or mastery testing where students are classified as being masters or non-masters based on their performance. Note that these assessment practices are different from grading practices that may include additional cognitive and behavioral factors to represent student achievement (see Brookhart, 2011). Mastery testing is analogous, but not identical, to the decisions that are made in credentialing programs. Specifically, credentialing examinations attempt to maximize measurement information at the point of the target candidate (e.g., minimally competent, minimally qualified). Information about candidates who are far above or below this point might be interesting but is not critical to the decision that the examination is intended to support.

More often, education assessments are designed to measure a greater depth and breadth of student achievement given the intended uses. The public is generally more familiar with information reported from these types of testing programs because they may have experienced it as a participant or read about it in popular media. Within the PreK-12 education context, these experiences may have included statewide or provincial student assessments, commercially available achievement tests (e.g., Iowa Tests of Basic Skills, Advanced Placement), higher education admissions (e.g., ACT, SAT, GMAT, MCAT). In each of these instances, a wide distribution of measurement and, ultimately, scores are expected to account for the range of abilities in the student population and to better differentiate across this range for education purposes (e.g., instruction, policy). This expectation of broad measurement holds whether scores are reported as multiple classifications of student achievement (e.g., Basic, Proficient, Advanced) or interpreted relative to norms (e.g., percentile ranks, normal curve equivalent scale scores).

Measurement of the depth and breadth of coverage of an educational program is generally more aligned with expectations for outcomes evidence in program accreditation for many professions. However, because program accreditation standards may not be prescriptive about the breadth and depth of measurement (or understand the difference in the focus of measurement), institutions will often seek to use the results of credentialing programs to serve as a proxy for more developed outcomes evidence. This mismatch represents an opportunity for all parties (i.e., credentialing programs, education and training programs, program accreditation bodies) to better define and develop the evidence that each entity needs to support its intended interpretations and uses.

Construct-relevant accommodations

“Fairness” as defined by the Test Standards includes multiple elements (e.g., test design, development, administration, validity of test score interpretations, accommodations). The focus of this section is about accessibility as it relates to the purpose of the examination and the intended use of the resultant scores and decisions. Examinations are intended to be a representative sample from a domain of interest and yield scores that are reflective of an individual’s ability in the domain. From these scores, we draw conclusions about whether these abilities translate to practice. Therefore, an important part of the validity argument is the extent to which we can make these claims. Operating in situations where cognitive and/or psychomotor abilities are measured, it is challenging to define, collect, and evaluate evidence that supports these claims. This challenge is heightened when responding to requests to accommodate or modify standardized conditions through which this information is collected and these inferences are made.

Conceptually, accommodations are designed to facilitate an individual’s demonstration of their ability while maintaining the meaning of the interpretation of a score or decision that results from that performance. Or to put it in the language of the Test Standards, accommodations are intended to remove construct-irrelevant barriers that could threaten the validity of the score interpretation. The differences between education and credentialing programs can be particularly challenging because this is an area where psychometric theory and practice intersects with legislation and regulations. Within the United States, these laws and regulations are characterized through the Americans with Disabilities Act (ADA, 1990) and the Individuals with Disabilities Education Act (IDEA, 1990).

A core tenet of the IDEA is the development of an individualized education plan (IEP) for students that specifies the accommodations and modifications that are permitted based on the student’s disability(ies). When allowed, these accommodations are often implemented as extended time, providing readers or scribes, linguistic translation or adaptation, or assistive technologies like text to speech. Considerable research about students’ academic achievement and types of accommodations has been conducted to better understand the impact of accommodations (see National Center on Educational Outcomes, http://ncee.info). However, questions about what constitutes appropriate

---

Note that the law was originally passed as the Education for All Handicapped Children Act (EAHCA, 1975) and renamed as the Individuals with Disabilities Education Act when it was reauthorized in 1990.
accommodations in the learning environment persist. For example, has a student receiving a reader or text to speech accommodation demonstrated the ability to read (e.g., decoding, using context clues, comprehension)? Moving beyond the learning environment, these questions extend to requests for accommodations in credentialing settings.

The Test Standards make a distinction between an accommodation and a modification that would change measurement of the construct and the interpretation of the performance (AERA, APA, & NCME, 2014, pp. 190-191). The goal of implementing accommodations is to remove construct-irrelevant barriers. However, what about construct-relevant barriers? This distinction is particularly important for credentialing examinations where job-related characteristics need to be distinguished from market-influenced characteristics. An attorney who needs extended time to prepare a memo or brief may be substantively competent but not ultimately survive in business. Conversely, a healthcare provider operating in a time sensitive environment may not have the luxury of unlimited time to demonstrate competency or meet a minimum market-based expectation of performance. Therefore, the interpretation of a job-related accommodation is clearly domain-specific. However, there is a grey area in the interpretation of what constitutes “reasonable” accommodations as required by law.

In practice, the opportunity to define the construct and what can be considered a reasonable accommodation is at the program design or redesign level and further articulated during the job analysis phase. This permits consideration of what can be considered job-related or not at the domain-specification and test development phases. In addition to the laws and regulations that influence accommodation practice, caselaw and amendments to the legislation has rendered this an area in testing that continues to evolve. Readers seeking additional information are encouraged to look for guidance from published resources (see Jacobs, 2016; Balasa, 2014) and to consult with their attorney.

Conclusions and recommendations
This article focused on some important distinctions between the application of test development and validation practices in the education and credentialing sectors. It is important to remember, though, that there are also considerable similarities in professional expectations. The Test Standards discuss validity, reliability, and fairness as applicable across testing programs, with specific differences highlighted in narrative explanations or in chapters unique to the setting. These distinctions will be driven by the intended interpretation and use of scores for the program.

As noted earlier, one of these important differences is the extent to which evidence of alignment among curriculum, instruction, and assessment is a part of the program. For education applications, this evidence is essential, whereas for credentialing programs there is an expectation that the influence of practice will outweigh the perspective of education or training programs. However, the inclusion of education perspectives in the domain specification process (e.g., job analysis) can be valuable as it can often provide advanced notice of declining or emerging practices that may be observed among practitioners in the future.

The focus of measurement for credentialing and education testing programs is usually different in terms of both the breadth and depth of content coverage and the range of performance that may be desired for interpretation and reporting. As such, it is important to avoid the intuitive comparisons of examinations developed and administered in these different environments. It is also advisable to overcome the temptation to co-opt scores or decisions from credentialing programs for purposes that they are not intended to support. Specifically, these sources of information provide limited (at best) support for program evaluation. Yet as program accreditation entities expect better outcomes evidence, these organizations as well as education and training programs may unknowingly push to misuse credentialing information. This practice presents an opportunity for assessment literacy regarding appropriate and inappropriate uses of scores from credentialing programs.

The intersection of legal and psychometric considerations in when and how to allow accommodations can be contentious, particularly as new assistive technologies further reduce barriers that may have existed in prior generations. Yet the reasonableness of the accommodation may be differentially observed in education programs when compared to the practice of a given profession. In addition to guidance from the Test Standards, the consideration of acceptable accommodations cannot be disentangled from the job-related context and construct. Programs and practitioners are encouraged to have conversations about appropriate accommodations during key phases of test development and validation, with guidance from legal counsel regarding how to navigate potentially competing strategies about how to respond to the challenge.
References


Debra P. v. Turlington, 474 F. Supp. 244 (M.D. Fla. 1979), aff’d in part, rev’d in part, 644 F.2d 397 (5th Cir. 1981); on remand, 564 F. Supp. 177 (M.D. Fla. 1983), aff’d, 730 F.2d 1405 (11th Cir. 1984).


Our Journey Continues as their Journey Begins...