

CLEAR Exam Review

Volume XIX, Number 2
Summer 2008

A Journal

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

Editing and composition of this journal have been written 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 Stephanie Thompson at (859) 269-1802, or at her e-mail address, sthompson@clearhq.org, for membership and subscription information.

Advertisements and Classified (e.g., position vacancies) for CER may be reserved by contacting Janet Horne 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

Janet Ciuccio
American Psychological Association

Steven Nettles
Applied Measurement Professionals

Jim Zukowski
Applied Measurement Professionals

Coeditor

Michael Rosenfeld, Ph.D.
Educational Testing Service
Princeton, NJ 08541-0001
mrosenfeld@ets.org

Coeditor

F. Jay Breyer, Ph.D.
Prometric
2000 Lenox Drive
Lawrenceville, NJ 08648
jay.breyer@prometric.com

CLEAR Exam Review

VOLUME XIX, NUMBER 2

SUMMER 2008

Contents

FROM THE EDITORS 1

F. Jay Breyer, Ph.D.
Michael Rosenfeld, Ph.D.

COLUMNS

Abstracts and Updates 2
George T. Gray, Ed.D.

Technology and Testing 7
Robert Shaw, Jr., Ph.D.

Legal Beat 11
Dale J. Atkinson, Esq.

ARTICLES

Addressing Nonresponse in Surveys 14
Anne Wendt, Ph.D., RN, CAE

The Design of Innovative Item Types: Targeting Constructs, Selecting Innovations, and Refining Prototypes 18
Cynthia G. Parshall, Ph.D. and J. Christine Harmes, Ph.D.

Evidence-Centered Design: A Lens Through Which the Process of Job Analysis May Be Focused to Guide the Development of Knowledge-Based Test Content Specifications 26
Richard J. Tannenbaum, Stacy L. Robustelli and Patricia A. Baron

From the Editors

Welcome to the Summer 2008 issue of the CLEAR *Exam Review*. We have three columns and three articles that we think you will find interesting and informative.

George Gray, in “Abstracts and Updates,” describes a new book “*A Rasch Primer: The Measurement Theory of George Rasch*” focusing on item response theory (IRT). He also discusses a number of articles dealing with the following topics: constructed-response items, standard setting, and validity.

Robert Shaw, in “Technology and Testing,” continues with the topic of biometric security systems. He describes weaknesses of biometric systems, biometric system costs, and privacy issues. This is the second of a two-part series dealing with biometric security systems.

Dale Atkinson’s legal column discusses a case involving a situation in which a licensure applicant admitted on the license application to having a psychological impairment. The licensing board requested the applicant undergo and pay for a psychological evaluation. The applicant brought suit and the resulting legal issues and outcomes will be of interest to any board dealing with evaluating license applications.

This issue also contains three feature articles. One is written by Anne Wendt and addresses the issue related to nonresponse in survey work – specifically related to practice analyses. The second article by Cynthia Parshall and Christine Harmes deals with the importance of matching the intent and purpose of measuring examinees with the selection and use of innovative items types. The authors also discuss various ways to refine the design and format of these kinds of items. This article provides a listing of suggested steps for evaluation and implementation if you are considering the use of some of these item types. The third article is written by Richard Tannenbaum, Stacy Robustelli, and Patricia Baron and discusses the incorporation of evidence-centered design aspects to help guide the job analysis process and the production of test content specifications. This article discusses the often asked – but rarely answered – question of how to move from job analyses to test specifications and it gives some suggestions to accomplish that goal in a practical and efficient manner.

As always, if you have an idea for an article or topic you feel would interest CER readers, please send it to either of us at the addresses on the facing page.

Abstracts and Updates

GEORGE T. GRAY

George T. Gray, EdD is director of test development
Professional Development Services, ACT, Inc.

Updates covered in this issue deal with a new book about one-parameter (Rasch) measurement models. Also articles on topics such as constructed-response items, standard setting, and validity in theory and practice are reviewed.

Rasch Measurement Models

Mead, R. (2008) *A Rasch Primer: The Measurement Theory of Georg Rasch*. Data Recognition Corporation, Maple Grove, MN, 53 pp.

In a recent review of the IRT chapter of the Fourth edition of the weighty (literally and figuratively) handbook *Educational Measurement* that was released in 2007, Bert Green, professor emeritus at Johns Hopkins states that “Yen and Fitzpatrick explain item response theory (IRT) succinctly but thoroughly. Although the Rasch model is discussed, along with many others, Rasch practitioners will miss their special terms, like ‘infit’ and ‘outfit,’ since the technical details here are all in the three-parameter tradition.” (Green, 2008, p. 196). Mead’s (2008) monograph presents the other side of the coin, explaining why the Rasch model is different. He even distinguishes Rasch models from IRT, relegating the latter term to the two and three parameter models endorsed by other authors. He also explains why one would use a Rasch model exclusively rather than just using it when it fits the data best or when the sample size is too small to use a three-parameter model (the three parameters being item difficulty, item discrimination, and a pseudo-chance or guessing parameter).

For those who are not math whizzes and wish to obtain an understanding of Rasch measurement Wright and Stone’s (1979) volume *Best Test Design* is still hard to beat. It includes step-by-step examples and features some data manipulation that can be replicated with a calculator. Mead’s monograph is less “user friendly” in this sense, but it does offer greater benefits in a number of ways. First, it is updated with almost thirty years of additional work. Second, it covers a variety of Rasch models, not just the one that is used for dichotomous data. Third, the text is so well-written that it is fairly easy to follow many of the major points that are made without needing to spend a lot of time digging around in the Greek letters of the equations. Fourth, the monograph is available at no charge as a download from the Data Recognition Corp website (www.datarecognitioncorp.com).

Over 40 years ago, Wright offered the following perspective: “In all of these approaches to item analysis...at least two parameters are sought for each item. Attempts are made to estimate not only an item difficulty, the response ogive’s horizontal intercept at probability one-half, but also an item discrimination, the ogive’s slope at this intercept. Unfortunately, this seemingly reasonable elaboration of the problem introduces an insurmountable difficulty into applying these ideas in practice. There has been a run-

ning debate...as to whether or not there is any useful way by which some kind of estimates of item parameters like item discrimination and item 'guessing' can be estimated....(D)ichotomous response data available for item analysis can only support the estimation of item difficulty..and attempts to estimate other item parameters are necessarily doomed." (Wright and Stone, p. ix)

While Mead takes the view that Rasch models are clearly preferable to two or three parameter IRT, he makes the case with more subtle persuasion than Wright did many years ago. Mead states that "the most popular motivation for using Rasch's models is that they are extraordinarily easy to use compared to the Item Response Models (IRT)." (p. 2) Mead states that he has attempted "to draw as sharply as possible, the distinction between Rasch measurement and the more complex IRT models, without criticizing IRT directly. In contrast to IRT, Rasch's interest was how best to extract all information from the data relevant to the construct of interest, not how to reproduce the data most precisely." (p. 2)

In addition to the most familiar Rasch model that includes dichotomous right/wrong scoring, Mead presents the partial credit and rating scale models. Those who are involved with judges as well as examinees and items will be interested in the multi-faceted Rasch model.

One often hears the term "objective measurement" associated with Rasch models. Mead explains that "Rasch chose the term *specific objectivity* to characterize his new models: *objective* because they allow comparisons between items without reference to the people and comparisons between people without reference to the items; *specific* to distinguish it from all other uses of objective but also to emphasize that this property is not demonstrated once and for all for all potential situations." (p. 9)

Rasch model advocates will continue to analyze data using Rasch models exclusively, but this monograph may encourage others to use Rasch models more extensively, not just for applications such as dichotomously scored items that have small sample sizes.

Constructed-response items

Hogan, T.P. and Murphy, G. (2007) Recommendations for preparing and scoring constructed-response items: what the experts say. *Applied Measurement in Education* 20(4), 427-441.

One usually expects to find mathematically intensive arti-

cles in *Applied Measurement in Education*. This article is a refreshing low-tech literature review that seeks to find consensus of expert opinion to guide practice. The authors analyzed recommendations for preparing and scoring constructed-response items (items for which the examinee must supply rather than select the correct response) for 25 measurement textbooks and chapters. This approach enabled the authors to seek consensus on recommendations in the literature and to review any empirical support for the recommendations.

Of twelve recommendations compiled for preparing constructed-response items, four were listed in more than half of the references:

1. Cover logistics, time limit, age limit, point value.
2. Avoid using optional items.
3. Define question/task clearly.
4. Relate to instructional objective(s), test blueprint.

Five recommendations for scoring the items were made in over half of the sources:

1. Score anonymously.
2. Score one item at a time.
3. Use a scoring rubric or ideal answer.
4. Separate mechanics from knowledge.
5. Use second reader.

A number of other recommendations for preparing and scoring items were endorsed by a smaller number of authors. In one case, the literature offered conflicting recommendations: use more items with less time per items and the completely opposite recommendation. This type of survey is an important starting point for developing a knowledge base for development and scoring of constructed-response items. Empirical studies would be a welcome next step.

Standard Setting

Yin, P. and Scoring, J. (2008). Estimating standard errors of cut scores for item rating and Mapmark procedures. *Educational and Psychological Measurement* 68 (1), 25-41.

This group of studies compared standard errors of cut scores using an Angoff item rating procedure and the Mapmark method, a method that combines the Bookmark method with item mapping. Studies were conducted using generalizability and decision study designs. The authors found that the cut scores were "generally consistent" for both the Angoff and Mapmark methods. They also noted that "it is clear that there is no one

standard error associated with a certain cut score.” (p. 25) The authors explain that “there are numerous possible standard errors corresponding to different universes of generalization and study designs. The magnitude of standard errors can be large or small depending on how restricted or relaxed the universe of generalization is and how many facets are considered in the study design.” (p. 40)

MacCann, R.G. (2008). A modification to Angoff and Bookmarking cut scores to account for the imperfect reliability of test scores. *Educational and Psychological Measurement* 68 (2), 197-214.

The author’s premise is that “...there is an ambiguity about the percentage of examinees identified as being below the level of minimal competence. For both methods, it will be shown that this percentage varies as a function of the length of a test, and given that the length of a test often depends on arbitrary factors, then the associated percentage in the ‘failure’ band becomes ambiguous.” (p. 198) He cites a hypothetical example in which as test length increases, along with increasing measurement precision and reliability of test scores, the percentage of examinees identified as less than minimally competent declines. This is explained as the difference between the candidates observed scores and their underlying true scores. As the test becomes longer and measurement error is reduced, candidates drop out of the failing group whose observed scores incorrectly lead them to be incorrectly classified as failing. The remaining candidates are increasingly limited to those who have underlying true scores in the failing classification. The distinction is made between the observed performance on a given day and underlying ability that would be measured over different occasions and a larger number of items. MacCann proposes a formula for adjusting cut scores for unreliability.

Plake, .B.S. (2008) Standard setters: Stand up and take a stand! *Educational Measurement: Issues and Practice* 27(1), 3-9.

This is the published version of Plake’s 2006 Career Award Address at the annual meeting of the National Council on Measurement in Education in 2007. As excerpts have been cited in a previous version of this column, no further details will be provided here: suffice to say that the article addresses a number of practical issues and is well worth reading.

Validity in Theory and Practice

Talento-Miller, E. and Rudner L.M. (2008). The validity of Graduate Management Admission Test Scores: a summary of studies conducted from 1997 to 2004. *Educational and Psychological Measurement* 68 (1), 129-138.

For CLEAR readers, the value of this article comes more from the method used to conduct validity studies than the results of the studies themselves. 273 studies involving over 41,000 students are summarized in the article. Predictive validity is the concept of interest for the GMAT. Undergraduate grades and GMAT scores are used as predictor variables and first year or mid-program grades are used as the criterion variable. “In this study, simple correlations are used to indicate the relationship between individual predictors. Multiple regression analyses are used to calculate validity coefficients for combinations of predictors.” (p. 131) As the studies were limited to only those individuals who were admitted to programs, enrolled and completed their first year, the validity coefficients were adjusted for restricted range of the data. GMAT total score was found to be the best single predictor of program grade point average.

Cizek, G., Rosenberg, S.L., and Koons, H.H. (2008) Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement* 68 (3), 397-412.

This provocative study is based on an analysis of data readily at hand. The authors reviewed information on tests in the 16th edition of the *Mental Measurements Yearbook*, focusing on four questions:

1. To what extent do validity reports conform to the major aspects of modern validity theory?
2. What specific sources of validity evidence are typically reported?
3. Do the sources of validity evidence gathered and reported for various kinds of tests differ?
4. What validity factors are considered most important? (p. 399)

The authors acknowledge the limitations of using a second hand source for information. In addition, the sample was expected to be weighted toward tests where “the producer believed that sufficient technical evidence was available for the test to be submitted for review.” Validity characteristics coded included (a) whether the review presented a unitary perspective on validity evidence, (b) whether the review cited a contemporary validity reference, and (c) whether the review referred to validity as a characteristic of a test or as a characteristic of scores or inferences. (p. 401)

A total of 283 tests were included in the study, covering a broad range of types of tests. Of the total sample, 54 tests were achievement tests. The results of the initial breakdown of data were striking. 2.5% of the tests took a unitary perspective on validity. Only 9.5% cited the *Standards for Educational and Psychological Testing* or Messick's chapter on validity in the 3rd edition of *Educational Measurement*. Approximately a quarter of the publishers treated the conception of validity as a "characteristic of test score, inference, or interpretation" in keeping with current concepts. 30% of the references treated validity as a characteristic of the test, and in almost half of the cases, "no clear indication of validity perspective could be discerned." (p. 404) The most frequent types of validity evidence reported were construct (58% of tests), concurrent criterion-related (51%) and content (48%). (p. 406)

The authors were favorably impressed with the *Mental Measurements Yearbook* as a data source. They also found encouraging the fact that test publishers typically did provide validity evidence (typically two sources). Among the less encouraging findings was a lack of regard for (or rejection of) unitary validity theory as it appears in the *Standards* or the chapter on validity from the 3rd edition of *Educational Measurement* which is cited in the *Standards*. "Of equal interest is the fact that there appears to be a lingering (mis)perception of validity as adhering to a test and that validity is represented as comprising various kinds. These perspectives stand in contrast to the fact that the unitary, inference-based view has been acknowledged—at least in theory—for nearly 20 years." (p. 409)

Lissitz, R.W. and Samuelsen, K. (2007) A suggested change in terminology and emphasis regarding validity in education. *Educational Researcher* 36 (8), 437-448.

Embretson, S. (2007) Construct validity: a universal validity system or just another test evaluation procedure? *Educational Researcher* 36 (8), 449-455.

Gorin, J.S. (2007) Reconsidering issues in validity theory. *Educational Researcher* 36 (8), 456-462.

Mislevy, R.J. (2007) Validity by design. *Educational Researcher* 36 (8), 463-469.

Moss, P.A. (2007) Reconstructing validity. *Educational Researcher* 36 (8), 470-476.

Sireci, S.G. (2007) On validity theory and test validation. *Educational Researcher* 36 (8), 477-481.

Lissitz, R.W. and Samuelsen, K. (2007) Further clarification regarding validity in education. *Educational Researcher* 36 (8), 482-484.

Kane, M.T. (2008) Terminology, emphasis, and utility in validation. *Educational Researcher* 37 (2), 76-82.

This article and series of responses is guaranteed to get the reader immersed in thinking about current validity concepts. The background is presented at the beginning of Lissitz and Samuelsen's article. They list a number of concerns with Messick's unitary concept of validity that is conceptually based on what used to be known as construct validity. Lissitz and Samuelsen are primarily interested in content validity but they make it clear that they do not intend for content validity to be viewed as the basis for a unified validity concept. They do argue that "construct validity as it currently exists has little to offer to test construction in educational testing." (p. 440). They suggest that "the approach that is being developed should initially be focused on the content elements of the assessment (what we call *internal* and theoretical), their relationships, and the student behavior and cognitions that relate to those elements as they are being processed (e.g., cognitive theories of cognitive processes)." (p. 440)

Internal factors in their model are practical and theoretical. Practical concerns are divided into content and reliability issues. These include sources of evidence such as analysis of the curriculum, determining match between items and the test blueprint, bias and sensitivity review, and internal consistency reliability. The theoretical perspective is described as a latent process with sources of data such as item analysis, inter-item correlations, and factor analysis. External factors include two types of practical factors: utility and impact, as well as theoretical factors pertaining to a nomological network. Sources of evidence for the latter include statistical tools such as analysis of variance, confirmatory factor analysis or structural equation modeling.

Lissitz and Samuelsen's article is followed by a number of responses. Embretson's perspective is that "content validity is not up to the burden of defining what is measured by a test. In fact, I believe that relying on content validity evidence, as available in practice, to determine the meaning of educational tests could have a detrimental impact on test quality." (p. 449) She favors a universal system for construct validity.

Gorin's position is that "Lissitz and Samuelsen's conceptualization may be seen as a move backward toward methods that have been shown historically to be problematic for measurement theory and practice." (p. 456) "Historically, the use of content validity tools such as operational definitions as indicators of score meaning has been tried and discarded." (p. 457) She believes that construct based

validity theory advances methods, and cites the use of cognitive models of test constructs as an example.

Mislevy also disagrees with the content validity perspective and focuses on a design argument –use argument model. He prefers Embretson’s phrase “construct representation argumentation for construct validity” rather than using a term such as content validity.

Moss worries that “the conception of validity...that Lissitz and Samuelsen propose appears to move away from a generative understanding of validation as scientific inquiry reflected in the unitary approach, (back) toward a representation of validity in terms of general methodological prescriptions that the unitary approach was intended, at least in part, to overcome.” (p. 470)

Sireci’s comments are based on four conclusions derived from the validity literature:

- Validity is not a property of a test. Rather it refers to the use of a test for a specific purpose.
- To evaluate the utility and appropriateness of a test for a particular purpose requires multiple sources of evidence.
- If the use of a test is to be defensible for a particular purpose, sufficient evidence must be put forward to defend the use of the test for that purpose.
- Evaluating test validity is not a static, one-time event; it is a continuous process. (p. 477)

Although not in agreement with Lissitz and Samuelsen’s model for validity, Sireci does indicate a sympathy for the perspective that construct validity as a unitary model is not completely satisfying. “Paramount among (its imperfections) is that it is extremely difficult to describe to lay audiences...It is hard to describe an underlying latent variable...It is far easier to talk about the content domain measured...Thus the concept of content validity is much more palatable and understandable than the concept of construct validity to policy makers and the general public.” (p. 478)

With the scorecard of expert opinion being heavily weighted against them, Lissitz and Samuelsen were provided 2500 words or less to respond to the five reviews of their paper. It would be difficult to select a highlight or two from their response. In the end, they stick to their major points and acknowledge that a change in terminology will be difficult. This will certainly be true in the measure-

ment community, but anyone who has labored in the field of test development will acknowledge that the content validity concept has never been set aside by the general public. It resonated with the lay community decades ago and has never gone away.

Kane’s commentary appears in a later issue of *Educational Researcher*. He concludes that the Lissitz and Samuelsen approach would entail a much more narrow focus for validation with no justification of the use of test scores. He states that, “I think if Lissitz and Samuelsen’s proposal were adopted, it would tend to cause confusion because, without achieving any substantial advance or solving any major problem, it abandons assumptions and terminology that have evolved for more than 50 years.” (p. 81)

References Cited

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Psychological Association.
- Brennen, R.L. (Ed.) (2006) *Educational Measurement* (4th ed.) Westport, CT: Praeger Publishers.
- Green, B. (2008). Book review: *Educational Measurement* (4th Ed.) edited by R.L. Brennan *Journal of Educational Measurement* 45(2) 195-200.
- Messick, S. (1989). Validity. In R.L. Linn (Ed.) *Educational Measurement* (3rd Ed), New York: Macmillan.
- Spies, R.A. and Plake, B.S. (Eds) (2005). *The sixteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurement.
- Wright, B.D. and Stone, R.L. (1979) *Best Test Design: Rasch Measurement*. Chicago, IL: MESA Press.

Technology and Testing

A Continuation of Biometric Security Systems Review

ROBERT C. SHAW, JR., PHD

Robert C. Shaw, Jr., PhD, is a Program Director for Applied Measurement Professionals, Inc.

This is the second part of a biometric security systems review. The Winter 2007 article described three methods by which persons' identities are commonly authenticated: (1) things people memorize, (2) objects people carry, and (3) measurements of unique human traits. The biometric part of a security system will include a sensor, an information extractor, a template database, and template matcher, which implies that any of these systems requires an enrollment phase followed by a matching phase. The utility of a particular biometric like a fingerprinting or facial geometry system is linked to several characteristics that affect system accuracy, how well people accept collection of the biometric, and how strongly the biometric resists circumvention. Security systems that incorporate biometrics are applied in one of two broad schemes: (1) to positively verify someone who should be able to access a system and (2) to negatively identify a bad actor who should be denied access. The first article on biometrics concluded by acknowledging that these systems can erroneously miss or make matches that are false negative and false positive results. Additional topics associated with the accuracy of biometrics are addressed in this article along with a discussion of privacy issues.

Attacks on Biometric Systems

While biometric authentication systems are often strong, particularly compared to things a person can remember or carry, they can be fooled. Uladag and Jain have challenged fingerprint recognition systems and identified eight ways to attack them. These eight techniques are summarized as follows:

1. Present a fake biometric to the sensor.
2. Present an intercepted biometric to the sensor.
3. Compromise the extractor to produce a template chosen by the attacker.
4. Replace a genuine template in the extractor with a template chosen by the attacker.
5. Add a fake template, modify an existing template, or remove a template from the database.
6. Modify the matcher to yield an artificially high matching score when a new template is presented.
7. Manipulate the transmission of a template from the database to the matcher.
8. Overwrite the result from the matcher.

One must think like a criminal to build systems that will be difficult to expose. While I was initially hesitant to spell out these attacking techniques, it seemed best to share them. After all, the criminals are aware of them. Attacks should be expected, since they are not extremely difficult to pull off. For example, Uladag and Jain attacked 11 finger-

print systems with a fake silicone finger and succeeded no less than two-thirds of the time after making multiple attempts on each system. A troubling outcome was that each of these 11 systems accepted the fake finger at least some of the time.

In another experiment, Uladag and Jain attacked a database of 160 accounts using a computer program. The false positive rate for a match was set at 0.1%, indicating that 1 in 1,000 attacks should have been accepted on average. Attempts to compromise each of the 160 accounts continued until successfully breached. The range for the frequency of attempts before a successful account breach was 132 to 871 with a mean of 271 indicating that a typical computer attack succeeded in far fewer than 1,000 attempts.

Uladag and Jain pointed out that a system administrator could set a stopping rule after too many false matches had occurred on a particular day, anticipating that such an occurrence was likely associated with an attack. However, a patient attacker with enough time could still succeed in a relatively short time. For example, 1,000 attacks could be metered out at 20 per day over 50 days and still fall short of an administrator's stopping rule.

Biometric System Costs

Rosenzweig, Kochems, & Schwartz (2004) summarized costs for the following sample of biometric systems in Table 1.

Biometric systems are typically classified into physical and behavioral categories based on the feature a system measures (Olsson 2003; Podio & Dunn). Measurements of physical features, like those described in Table 1, tend to produce more accurate standalone systems. However, the price for accuracy is typically greater expense and complexity.

System	\$ for Hardware	Comments
Iris recognition	2,000	<ul style="list-style-type: none"> • A comprehensive system will include other hardware, software, and typically licensing fees, so the whole system will cost much more than \$2,000
Hand geometry	2,000 to 4,000	<ul style="list-style-type: none"> • Considered a mature technology with more than a 30-year history • Only acceptable for positive verification applications
Print reader for one finger	1,000 to 3,000	<ul style="list-style-type: none"> • Expect a software licensing fee of \$4/user • Typical annual maintenance runs 15%-18% of the purchase price
Print reader for 10 fingers	25,000	<ul style="list-style-type: none"> • Typical annual maintenance runs 14% of the purchase price
Facial recognition	15,000	<ul style="list-style-type: none"> • Software licenses range from \$650 to \$4,500 • Closed-circuit television equipment increases the expense depending on the number of access points • The only system routinely used for covert enrollment

Behavioral measurements based on patterns observed while speaking, walking, or typing are typically less accurate than physical measurements in positive verification systems. Physical biometrics typically have error rates down around 1% or less while error rates as high as 10% to 20% are observed for systems based on measured behaviors (Jain, Ross, & Prabhakar 2004). Monroe & Rubin (2000) emphasize that a behavioral-based system like keystroke recognition tends to be added to other authentication procedures rather than used as a stand-alone, primary source of positive verification. However, in considering traditional test enrollment systems based on things candidates can remember or carry, adding a behavioral biometric could strengthen the whole system for less expense than a physical biometric.

An administrator could apply keystroke recognition for static or continuous positive verification (Monroe & Rubin 2000). Keystroke recognition could be applied in a testing environment during a candidate's login, which is an example of static verification. However, this would not prevent a confederate from replacing the candidate after he

or she logged in. A test proctor should detect a candidate switch, but the proctor may not pay close attention or even take a bribe to ignore the switch. Continuous verification of keystroke samples taken throughout the period a candidate was logged into the system would create a stronger system. Authentication samples likely should be taken covertly in such a hypothetical system to deter cheating. However, continuous verification would be difficult to apply to a multiple-choice examination in which keystrokes are limited or intermittent verification obvious to the test taker.

Mitigating Privacy Threats

Concerns about privacy invasions associated with biometric systems were discussed by Rosenzweig, Kochems, & Schwartz (2004) and Jain, Ross, & Prabhakar (2004). For example, a system based on retinal patterns could be used to detect pathologies associated with diabetes and high blood pressure. DNA samples could yield information about the likelihood of genetic disorders in the person from which the sample was taken or in his or her children. An employee's concern about his or her privacy is easy to anticipate when insurability could be affected by information collected by a biometric system. Extending concerns about privacy to credentialing programs, candidates could justifiably become apprehensive that the owner of some types of biometric samples could give health risk ratings along with a competency determination to potential or current employers.

Most people would consider these scenarios as examples of abuse. Rosenzweig, Kochems, & Schwartz (2004) emphasize that concerns about abuse are based on fears that unwanted information will be collected without permission, used for purposes beyond the one for which it was first gathered, distributed to other entities without permission, and used to develop a more complete image of a person for surveillance and social control purposes. A program planning to add a biometric to its authentication system should anticipate these concerns and develop procedures that minimize or eliminate opportunities for abuse.

Rosenzweig, Kochems, & Schwartz (2004) offer a set of principles for systems that should be perceived as less threatening by enrollees:

- Enrollment should be overt; each enrollee should give his or her consent
- Positive verification presents fewer potential threats to privacy

- Because central storage could be subject to unwanted new uses, local storage of biometric templates is preferred
- Biometric templates should be the only residual of enrollment that are stored; raw information that could be visually recognized (i.e., images of faces) should not be stored
- Strong audit and oversight programs should be in place; biometric systems should be thoroughly and repeatedly tested
- A secondary recognition system should be in place for occasions when the biometric system fails or the result is inconclusive

Building a System

Just as we seek multiple pieces of evidence to support valid inferences from examination results, authentication systems with multiple features are best. This is true whether a biometric is a component of the system or not. Adding a biometric to existing authentication systems should increase accuracy. The more independent each feature is from others in the system, the more accuracy improves (Jain, Ross, & Prabhakar 2004). For example, a system based on speech and face data is likely better than one based on prints from two fingers. However, the more independent each measure is, the more expensive the system becomes since a different sensor is required for each biometric. In addition to increasing accuracy, adding biometric features to an existing system also provides a built-in backup or secondary authentication that can be applied when the primary fails.

Security can be enhanced by requiring presentation of multiple biometrics in a prescribed sequence for which the pattern is occasionally varied. For example, enrollees could be required to present their right index fingerprints followed by left thumbprints this week. The next week, enrollees could be directed to present their left index fingerprint followed by right middle fingerprints.

I want to take this opportunity to emphasize advantages of storing biometric templates on smart cards (Rosenzweig, Kochems, & Schwartz 2004) that a user would carry with him or her. Assuming that someone could not access a system until he or she presented the smart card and a new biometric that matched the one on the card, the risk from lost cards should be minimal. If templates were stored only on smartcards, then users

should be comfortable that their information will be kept confidential since their templates are not centrally stored. Using smart card biometric matching in conjunction with something one can remember like a password or PIN would yield a three-featured system that should be reasonably secure and easy to implement. It may sound as though there are just two features in this example, the biometric match and the password/ID number, but remember that simply possessing the card is part of the system. Only people who bring all three elements together would be authenticated.

Thinking through the scenario I just described would likely leave an administrator concerned that someone could tamper with the smartcard. Someone could steal a card and replace the template with that of an imposter who could present the compromised card and his or her biometric that is highly likely to match. The solution is to also store templates in a central database to which the template on the card would be compared. If that match was made, then the user could be asked to present a new template, which would be matched to the smartcard and central database. Such a system adds a fourth element to the system. Security would be increased by centrally storing templates, but the privacy threat also increases. There are no perfect solutions. The challenge lies in finding the balance.

In closing, I will point the interested reader to the Standards Council of Canada (<http://www.scc.ca/>) and the National Institute of Standards and Technology (<http://csrc.nist.gov/>) in the United States, which have established standards for authentication systems used in these two countries.

References

- Jain, A K, Ross, A, Prabhakar, S. 2004. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*. Vol. 14(1).
http://www.csee.wvu.edu/~ross/pubs/RossBioIntro_CSVT2004.pdf
- Monrose, F, Rubin, A D. 2000. Keystroke dynamics as a biometric for authentication. *Future Generation Computer Systems*. Vol 16. pp. 351-359.
<http://www.cs.jhu.edu/~fabian/papers/fgcs.pdf>
- Olsson, T. 2003. Strengthening authentication with biometric technology. The Information Security Reading Room. The SANS Institute.
<http://www.securitytechnet.com/resource/security/authen/1226.pdf>
- Podio, F L, Dunn, J S. Biometric Authentication Technology: From the Movies to Your Desktop
<http://www.itl.nist.gov/div893/biometrics/Biometricsfromthemovies.pdf>
- Rosenzweig, P, Kochems, A, Schwartz, A. 2004. Biometric technologies: Security, legal, and policy implications.
<http://www.heritage.org/Research/HomelandSecurity/lm12.cfm>
- Uludag, U, Jain, A K. Attacks on biometric systems: A case study in fingerprints.
http://biometrics.cse.msu.edu/Publications/SecureBiometrics/UludagJain_BiometricAttacks_SPIE04.pdf

Legal Beat

Not Just Another Test

DALE J. ATKINSON, ESQ.

He is a partner in the law firm of Atkinson & Atkinson.

<http://www.lawyers.com/atkinson&atkinson/>

In the interest of public protection, regulatory boards enforce the statutory requirements of assessing and licensing applicants to practice a particular profession. Licensing boards are generally empowered to determine whether applicants for licensure meet the qualifications for licensure set forth in the practice act and corresponding regulations/rules promulgated by the respective board. In addition to requirements related to applications, fees, personal history/good moral character assessments, and, perhaps, criminal background checks, most practice acts contain an education requirement, perhaps an experience requirement, and an examination requirement.

While previous *CLEAR Exam Review* articles focused on minimum competence examinations developed upon a practice analysis, assessment of the practice, blueprint, item development and statistical analysis of the performance of the exam items for ultimate use in the licensure process, this article will focus upon an entirely different form of examination. Not all evaluative measures are based upon subject matter knowledge of the profession. Certain additional evaluations related to the mental health of an applicant may be necessary and appropriate. The use of such an evaluative tool is based upon a grant of legal authority. Without the authority, regulatory boards will be unable to require an applicant to undergo a mental health evaluation as a prerequisite to licensure. Under these circumstances, many additional factors further complicate the application and decision-making process, including the expertise of the board and the confidentiality of the documents and eligibility determination.

Determining whether a board has the authority to require a mental health evaluation is a complex issue and likely involves an assessment of several areas of the law. Boards are encouraged to review and have some working knowledge of not only the practice act which is the guiding law, but of other ancillary statutes and regulations. These other laws may include disabilities legislation, confidentiality laws, open records and meetings requirements as well as many others. At times, federal legislation may also be relevant. All such applicable laws will determine what issues are relevant to the application process and, thus, help determine what questions to ask on the applications for licensure and renewals.

When assessing applications for initial licensure, and depending upon language within the practice act, boards may be authorized to consider matters beyond those specifically delineated in the qualifications for licensure section of the law. That is, the grounds for discipline section of the practice may (and should) read: "The board may refuse to issue, refuse to renew or may suspend, revoke...any individual...upon one or more of the following grounds... ." This empowering language may create opportunities for the board to ensure the public protection by assessing numerous factors beyond those

specifically set forth in the qualifications section of the law. For instance, the grounds for discipline section of the law may state that impairment or incapacity that prevents an individual from engaging safely and effectively in the practice may provide the board with grounds for administrative action. Under the language above, impairment may also provide for grounds to refuse to issue or renew licensure.

While an argument can be made that impairment that results in grounds to remove a license provides sufficient grounds to deny an application, legislative language may limit what the board can ask of applicants on the application and, in turn require of the applicant as a prerequisite to licensure. Consider the following.

An individual with a mental disability applied for licensure as an attorney in Wisconsin. She disclosed on her application her disability, specifically that the Social Security Administration had certified her as disabled and that she qualified for and received benefits indicating that she was unable to pursue gainful employment. Her disability involved chronic depression and fatigue for which she received no treatment or medication. The applicant authorized the release of her medical records, but her doctors had not retained them. Based upon the self disclosure information, the Wisconsin Board of Bar Examiners (board), before it would rule on her eligibility for licensure, directed the applicant to undergo a psychological evaluation at her expense (\$2,000). The applicant refused to seek the evaluation but offered to provide affidavits from employers and professors as to her fitness to practice. The board rejected this suggestion.

Further complicating this matter was the fact that the applicant graduated from the University of Wisconsin Law School which, subject to a particular time period, allowed her to receive licensure through a “diploma privilege” (without having to pass the bar exam). The board refused to act on her application without the psychological evaluation and the window of opportunity to become licensed without passing the bar exam expired.

The applicant filed suit against the state, the board, and the board members in both their individual and official capacities alleging numerous constitutional violations involving discrimination, the Americans with Disabilities Act (ADA), the Vocational Rehabilitation Act (Rehab Act), and others. After certain procedural exercises resulting in the dismissal of numerous counts of her complaint, the board, facing the prospect of trial, agreed that if the appli-

cant reapplied for admission, she could retain her diploma privilege and would not be required to undergo a psychological examination. Because the applicant received all the relief that she sought regarding the claim for an injunction, the court dismissed the action as moot. The applicant appealed both the dismissal of her claims for injunctive relief as well as her claim for damages that had been dismissed earlier through procedural motions.

The applicant’s first claim on appeal involved allegations of constitutional violations by the board in its demand that she undergo a psychological evaluation and consent to the disclosure of her medical records. The applicant argued that such a request violated her rights under the Fourth Amendment as an unlawful search. The court of appeals quickly disposed of this argument citing case law that a psychological evaluation is not a “search.”

The argument that the board deprived the applicant of her property right and liberty interest in pursuing her chosen profession also failed. The court noted that the right to pursue one’s profession is most certainly subject to reasonable governmental regulation, that her receipt of disability benefits signifies that the government considers her unable to pursue gainful employment, and that it was reasonable for the board, under these circumstances, to require a psychological fitness evaluation before licensure. The court continued, noting that the applicant’s desire for a speedy and routine approval of her application did not establish a property interest.

The applicant also argued that board request for an evaluation violated her right to equal protection under the law. The court again rejected this argument stating that the board must only license those persons “whose record of conduct justifies the trust of clients, adversaries, courts and others with respect to the professional duties owed to them.” It held that given the mental health history, the board request for an evaluation was rationally related to its interest in ensuring that only competent persons be admitted to practice law in Wisconsin.

Turning its attention to the disability claims, the court examined her claim under the ADA. The court affirmed the dismissal of the board based upon state immunity which was not abrogated by the enactment of the ADA. Regarding the ADA claim against the board members in their individual capacities, the court affirmed the dismissal because the ADA authorizes suits only against public entities. Under the Rehab Act, the court also upheld the dismissal against the board because the applicant present-

ed no competent evidence that the board received federal assistance, as required under the act. Finally, the court upheld the dismissal of the action against the individual defendants in their personal capacities because, again, the Rehab Act only authorizes suits against public entities. Because the board offered the applicant all the relief to which she may be entitled, and she cannot, as a pro se litigant, represent others, the court noted that the litigation was subject to dismissal.

In this case, the board ultimately determined that the applicant would not have to undergo a psychological evaluation. However, it appears from the opinion that a mandate of such an examination, under certain circumstances, may be warranted and would be within the purview of a board of bar examiners. Regulatory boards are encouraged to understand the bounds of their authority and, where necessary, to use such authorization to ensure applicants for licensure meet the criteria for licensure. The bounds of board authority may be vested in multiple sections of the practice act. While not suggesting that boards unfairly deny applicants licensure, a denial of licensure is subject to differing due process requirements than the removal of a license once granted.

Brewer v. Wisconsin Board of Bar Examiners, 2008 WL 687315 (7th Cir. 2008)

Addressing Nonresponse in Surveys

ANNE WENDT, PhD, RN, CAE

National Council of State Boards of Nursing

Introduction

Many countries use surveys to measure characteristics of their population such as socioeconomic status or health. While the use of surveys may have worked well in the past, more recently there has been a decline in survey response rates (Groves, Dillman, Eltinge, Little, 2002). A decrease in response rates can affect the ability of the survey statistics to accurately reflect the characteristics of the population which would indicate nonresponse bias. However, it should be noted that the lower response rates do not necessarily indicate nonresponse bias and higher response rates do not necessarily indicate no nonresponse bias. Regardless of the response rate, if the nonresponders are very different in the characteristics that the survey is measuring from the responders, there is bias (Groves, Fowler, Couper, Lepkowski, Singer, Tourangeau, 2004).

When conducting a survey, it is often impractical to survey the entire universe of potential respondents. When the universe is very large, it is preferable to randomly sample the population to get a smaller group that can be examined in great detail. For this to be successful, the sample has to be adequately large to produce results with a useful degree of precision and the sample has to be generally representative of the population. Despite a well-crafted sampling design, a sample of surveys can be dramatically influenced by systematic non-response bias. This occurs when there is a characteristic among the potential respondents that makes them less likely to respond and that characteristic is relevant to the topic that the survey is addressing (unit level nonresponse behavior). An additional nonresponse issue for researchers to consider is when respondents do not complete the entire survey or complete only selected questions (item level nonresponse). Item level nonresponse will not be addressed in this article, however, readers are referred to the references provided hereafter for more information on this issue.

Using a practice analysis survey of nurses, we can examine the issue of individual non-response further. If most hospitals required their nurses to sign agreements that prohibited the nurse from saying anything about the nature of the work that they perform in the hospital, then the nursing activities that are specific to working in a hospital could be dramatically underrepresented in the survey results. Given that hospitals are one of the most common work-settings; this could have a substantial impact on how well the practice analysis results would reflect the practice of nursing in the United States. This would be an example of systematic response bias.

On the other hand, respondents that are not working in the profession may be less likely to respond because they believe, and correctly so, that their responses are not relevant to the purpose of the survey. Similarly, those people who do not check their e-

mail, misplace their regular mail, or are just very busy also might not respond. Yet, if their reason for not responding is not systematically related to an important aspect of the survey, then it doesn't introduce any systematic bias into the results. Without surveying the entire population, it is impossible to know with 100% certainty if any systematic bias has been introduced, but in the absence of a logical rationale for such a bias, such biases are considered to be trivial or nonexistent. Of course, one could be less than rigorous in attempting to find such a logical rationale for a bias and could erroneously conclude that no bias existed. In order to determine if there is a nonresponse bias for recent practice analyses of nurses, the National Council of State Boards of Nursing (NCSBN) began conducting non-responder studies. These nonresponder studies address the issue of individual or unit nonresponse behavior but not the issue of item nonresponse when respondents do not answer all of the survey questions.

Background

In 2006, NCSBN began a web-based continuous RN practice analysis of newly licensed registered nurses (NCSBN, 2008). At the conclusion of the first year of data collection from July 2006 through June 2007, 116,985 nurses were asked to complete a practice analysis survey. There were 13,763 surveys "returned" due to incorrect or invalid e-mail addresses. Of the 103,249 invitations that reached recipients, 23,253 respondents submitted surveys for a return rate of 22.5%. In order to determine if there was systematic nonresponse bias, the nurses who did not respond were contacted by telephone.

Methodology

A random selection of 494 nonresponders was drawn from the sample of nurses who were emailed the survey during the previous six months. The information on the 494 nonresponders included the telephone number they listed when registering for the NCLEX examination. Interviewers then attempted to contact the nonresponders. There were 66 disconnected numbers; of the remaining 428 numbers dialed by the interviewers, 50 of them led to direct contact and participation in the survey. Once telephone contact was obtained, the nonresponders were asked a series of questions beginning with their reason for not responding and their length of time in practice. Once that nonresponders were engaged, the interviewers asked them to rate 10 activity statements from the *Report of the Finding of the 2006-2007 Continuous RN Practice Analysis Survey* using the same scale as the practice analysis survey

(NCSBN, 2008). The 10 activity statements were selected by subject matter experts to represent those activities most likely to be performed by newly licensed nurses and those activities least likely to be performed such as those activities performed in a specialized area of nursing practice.

Results

Nonresponse Reasons

Nonrespondents were asked the reason for not responding to the survey that was e-mailed to them. They were asked to choose from a list of prepared options. This list of options was based on initial phone interviews conducted during previous nonresponder studies where participants were asked for a reason for not answering the survey. These answers were combined and a list of the most frequent responses was created. This list includes:

1. Too Busy
2. Did not care
3. Do not like/trust surveys
4. Did not receive it
5. Other

As seen in Figure 1, 52% (26 of 50) of non-respondents stated that they never received the initial e-mail survey while 42% (21 of 50) chose the option of "other". Some of the reasons noted under "Other" include, "My husband threw it away," "I just forgot" and "It was too long." The remaining 6% (3 of 50) either did not like/mistrusted surveys or were too busy to answer the initial survey.

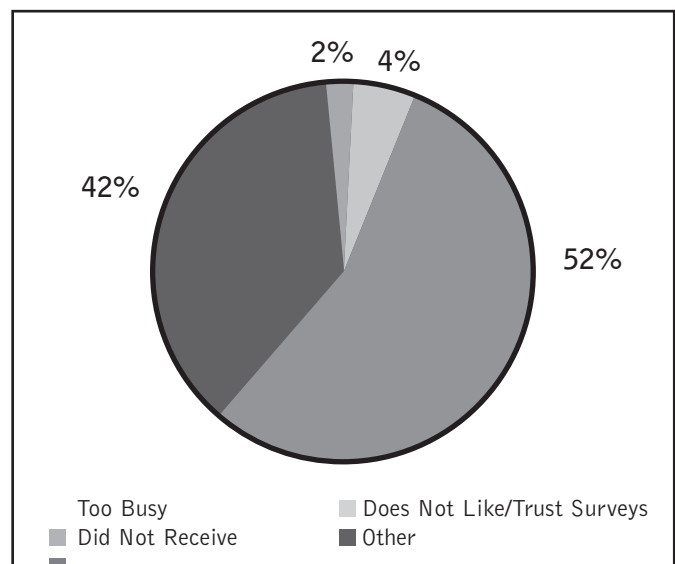


FIGURE 1. Reasons for Not Responding

Table 1. Importance of Activity Performance – Nonresponders vs. Responders

Apply Principles of infection control (e.g. hand hygiene, room assignment, isolation, aseptic/sterile technique, universal/standard precautions)		
	Importance	N
Nonrespondents	3.96	50
Respondents	3.94	17653
Administer and document medications given by common routes (e.g. oral, topical)		
	Importance	N
Nonrespondents	3.84	50
Respondents	3.83	17292
Participate in performance improvement/quality assurance process (formally collect data or participate on a team)		
	Importance	N
Nonrespondents	3.60	47
Respondents	2.94	3135
Perform emergency care procedures (e.g. cardio-pulmonary resuscitation, abdominal thrusts, respiratory support, automated external defibrillator)		
	Importance	N
Nonrespondents	3.57	44
Respondents	3.74	14025
Assess psychosocial, spiritual, cultural and occupational factors affecting care		
	Importance	N
Nonrespondents	3.76	49
Respondents	3.18	16581
Provide end of life care to clients and families		
	Importance	N
Nonrespondents	3.19	42
Respondents	3.38	3538
Supervise care provided by others (e.g. LPN/VN, assistive personnel, other RNs)		
	Importance	N
Nonrespondents	3.48	44
Respondents	3.23	16645
Serve as a resource person to other staff		
	Importance	N
Nonrespondents	3.61	41
Respondents	3.21	4144
Plan and/or participate in the education of individuals in the community (e.g. health fairs, school education, drug education, sexually transmitted diseases)		
	Importance	N
Nonrespondents	2.89	37
Respondents	2.71	11079
Teach clients and families about the safe use of equipment needed for healthcare		
	Importance	N
Nonrespondents	3.68	44
Respondents	3.35	4094

Average Months Working as an RN

Nonrespondents had been working an average of 13 months as an RN. Due to the time span between the initial survey and the nonresponder survey, participants in this study had been working longer than original respondents who had worked an average of five months.

Activity Statement Ratings

For each of the 10 activities, nonrespondents were asked to rate the overall importance of the activity considering client safety, and/or threat of complications or distress. They were asked to use the following scale: 1=Not Important, 2=Somewhat Important, 3=Important, and 4=Extremely Important. Table 1 shows the nonresponder importance ratings for each of the activity statements compared to the total group importance ratings of the survey respondents. As can be seen, importance ratings by nonrespondents were very similar to ratings by the original respondents. All ratings were within one point of one another.

Summary

Fifty nonrespondents from the 2006-2007 RN Continuous Practice Analysis Survey were called by interviewers. The majority of non-respondents did not remember receiving the initial survey. Nonrespondents had been working an average of 13 months as an RN as compared to the survey respondents who had been working an average of five months. Both cohorts generally agreed with regard to importance ratings of the activity statements. Using this data, it would appear that there may be no system-

atic differences in the responders versus nonresponders and the researcher could conclude that the statistics from the sample could generalize to the target population.

Survey researchers are often concerned about the nonresponse bias (Groves, Fowler, Couper, Lepkowski, 2004). The use of a nonresponder study can assist researchers to address the issue of nonresponse. NCSBN began using the methodology in 2006 and has been refining and enhancing the methodology to include additional nonresponders and additional activities in the telephone interviews.

References

- Dillman, D., Eltinge, J., Groves, R., and Little, R. (Eds.). (2002). *Survey nonresponse*. New Jersey: Wiley and Sons.
- Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., and Tourangeau, R., (2004). *Survey methodology*. New Jersey: Wiley and Sons.
- National Council of State Boards of Nursing. (2008). *Report of the finding of the 2006-2007 RN continuous practice analysis*. Chicago: Author.



The Design of Innovative Item Types:

Targeting Constructs, Selecting Innovations, and Refining Prototypes

CYNTHIA G. PARSHALL, PhD
Measurement Consultant

J. CHRISTINE HARMES, PhD
James Madison University

Introduction

The multiple choice item type has a well-established history of successful application, in both low- and high-stakes settings. The measurement community as a whole, over a period of decades, has developed substantial expertise in the use of this item type. This expertise is evident in the ready availability of consistent item writing guidelines, approaches to measuring higher level thinking, and statistical item analysis techniques.

Far more recently, many new item types have been developed and implemented in computer-based tests (CBTs; Parshall, Spray, Kalohn, & Davey, 2002). There is a great deal of excitement about these “innovative item types”, both within the measurement community and within some testing boards and user groups. However, the same overall level of expertise for the design and development of these new item types does not yet exist. While innovative items have been implemented successfully in many operational exam programs, it is also true that some exam programs have implemented innovative items without sufficient preparation, potentially leading to an unfortunate decrease in the quality of the exam.

For all these reasons, we suggest that when an exam program is considering the addition of innovative items, it is worth undertaking a thoughtful evaluation of the exam program’s construct needs, a thorough analysis of the challenges inherent in specific innovations, and a structured approach to the design of any new item types. We offer here a six-step design process that we believe can improve the quality and utility of innovative item types for any exam program. These steps are illustrated in Figure 1, highlighted in Table 1, and discussed in detail within this paper. We recommend that the steps be carried out, through multiple rounds of feedback and revision, prior to the operational development and administration of any new item type. If this thoughtful, careful approach is followed then the risks associated with new item types should be greatly minimized and the benefits offered should be substantially maximized.

Step 1 – Analyze the Exam Program’s Construct Needs

The first step in this model for innovative item type design involves the thoughtful consideration of an exam program’s specific goals and requirements. The test developers, along with subject matter experts (SMEs), should carefully consider those areas of the content that are currently being successfully addressed, as well as any that are not.

Any “missing pieces” in the current assessment should be acknowledged. The purpose of this step is to conduct a thorough analysis of the exam program in order to identify any deficiencies in the construct coverage that might be addressed through the use of innovative item types. This phase should result in a list of a few specific areas of content or cognition that ideally ought to be added to the exam program.

Step 2 – Select Specific Innovations

The next step in the design of item types is to consider the potential match between specific types of innovation and those construct areas that the exam program has identified as lacking. There are several aspects of innovation that may be useful for consideration: 1) assessment structure, 2) response action, 3) media inclusion, 4) interactivity, 5) complexity, 6) fidelity, and 7) scoring method (Parshall, Harmes, Davey & Pashley, in press).

These seven aspects of innovative assessments are not fully independent of one another; nevertheless, it can be useful to consider each element individually. Each element relates to important decisions that test developers must make when designing innovative items and their associated interfaces.

- *Assessment structure* defines the structure of the item presentation and the kind of response collected from the examinee. This term encompasses the more traditional description of “item type” or “item format”, while allowing for the greater breadth that may occur in some innovative assessments. Innovative assessment structures range from selected response items, through various forms of constructed response, through more elaborate multi-step tasks, to full-scale simulations.
- *Response action* refers to the means by which examinees provide their responses, along with the input devices used. The most common input devices are the keyboard and mouse. Other input devices, and other user actions, are less common but may be of high value in certain instances.
- *Media inclusion* covers the use in an item of media such as graphics, sound, animation, or video. In most instances, the media is provided as part of the item stem, but media may also be incorporated into response options. This type of innovation has broad application for many areas of content and construct.
- *Interactivity* describes the extent to which an item reacts

or responds to examinee input. This is likely to be minimal in discrete item formats, such as selected response items, while more elaborate assessment structures, such as simulations, may include extensive interactivity.

- *Complexity* refers to the number and variety of elements associated with an item that the examinee may need to interpret or use in order to provide a response. Greater complexity can be associated with contextualized assessments that target higher-order thinking. However, it is important to ensure that only content relevant complexity is included.
- *Fidelity* considers the degree to which an item provides a realistic and accurate representation of the components and tasks related to the construct being measured. An appropriate level of fidelity to target is often the level that is necessary to address the construct goal, but no higher. This is due to the costs that usually accompany increases in fidelity.
- Finally, *scoring method* addresses how examinee responses are translated into quantitative scores. Many innovative item types included in operational exam programs use dichotomous scoring, but various approaches to partial-credit scoring, as well as complex modeling algorithms, have also been used.

Each of these aspects of innovation can be considered in terms of its most likely costs and benefits (Harmes & Parshall, 2007; Parshall & Harmes, 2008). In general, the most promising benefit to any type of innovation is the potential to improve the measurement of the underlying construct. The most frequent increases in costs to the exam program occur as a result of custom programming needs and psychometric research and development requirements. Other examples of possible costs include production expenses when media will be included, or the need for extended training and tutorials when the new item type relies on higher levels of examinee computer skills. These developmental demands may all impose costs on personnel time and effort, typically resulting in increased financial costs.

An exam program can thus begin to determine what costs are likely to derive from the inclusion of specific innovative item types. For example, a program might decide to expand construct measurement by adding audio to the exam. In this instance, certain increased costs might result from a need for additional programming to enhance the test delivery software. Further costs might include the

audio production itself (e.g., voice actors, studio time, editing). On the other hand, benefits for this innovation might include an increase in the domain coverage of the exam, and a reduction in the dependence on examinees' reading skills (assuming a construct in which reading skills are not relevant).

The result of this step should be the selection of a few specific innovative item types for prototype development.

Step 3 – Design Initial Prototypes for Internal Discussion

The third step in this approach to item type design is to develop one or two prototype items of each selected innovation. Depending on the type of innovation and the construct needs, this step can require substantial cognitive work on the part of test developers and SMEs. Thoughtful consideration is needed even in relatively simple cases, such as changing traditional selected response items to free entry, adding graphics to item stems, or utilizing touch screens as the input device. More elaborate innovations, such as those that include multi-step situated tasks, interactivity, or complexity, warrant a thorough analysis of possible approaches and their implications. This cognitive analysis will often be time-consuming. However, it is also close to the heart of assessment and should not be neglected or short-changed. It is in this process that the basic goals for the item type are made practical and applicable to the exam program.

Once some basic decisions have been made regarding how the innovation should be used to improve measurement of the construct, then an initial prototype can be produced. For example, if the addition of video was identified as potentially useful for addressing a construct deficiency, then one or two sample video items should be developed. These examples should reflect the decisions made about any item type features or characteristics that are to be included. Prototype innovative items, or “mock-ups,” can be developed fairly easily in software applications such as PowerPoint. These prototypes do not need to be fully functional, but they should illustrate all the intended item type features. Even paper-and-pencil prototypes, along with textual descriptions of specific functionality or implementation details, can be very useful at this initial stage.

Once these prototypes are available, then the full set of test development stakeholders should thoroughly evalu-

ate each item type for its potential inclusion in the exam program. Test developers and SMEs should consider, first, whether a given prototype appears likely to address the needed content or cognitive element. However, this SME evaluation should expand beyond a traditional content validity review. It should also address the prototype item's potential value for and impact on additional aspects of the item type implementation, such as its navigation, required response actions, and type of interactivity.

Each prototype should also be evaluated by all other stakeholder departments and teams. This evaluation work can serve as a type of “feasibility review” for each potential item type. At a minimum, these evaluations should address: the usability of the item interface for the examinees, any technological aspects of item type development and delivery, item type scoring and psychometric needs, and preliminary cost estimates for each potential item type.

When feasibility challenges are identified, as they will be, team members should jointly seek potential solutions to these concerns. Assuming that the content review is positive, this step may be defined by various test development teams working together to answer the question, “Can we make it work?” Nevertheless, this review step will sometimes result in the elimination of one or more problematic prototypes from further consideration.

The remaining, more promising, prototypes are then revised based on these initial reviews. At this point, the prototypes are also more fully developed. In addition to the substantive revisions that may be needed, these revised prototypes usually differ from the initial prototypes in terms of their overall level of specificity as well as characteristics of the item type's appearance and functionality. These revised prototypes will be further developed and evaluated in the next step.

Step 4 – Iteratively Refine the Item Type Designs

This step in the design of innovative item types consists of three related sets of activities that are undertaken in a series of connected iterations (see Figure 1). Within each iteration, feedback is input and revisions are attempted, in order to arrive at an improved item type design.

The three sets of activities are generally conducted by different stakeholder groups. First, the content development

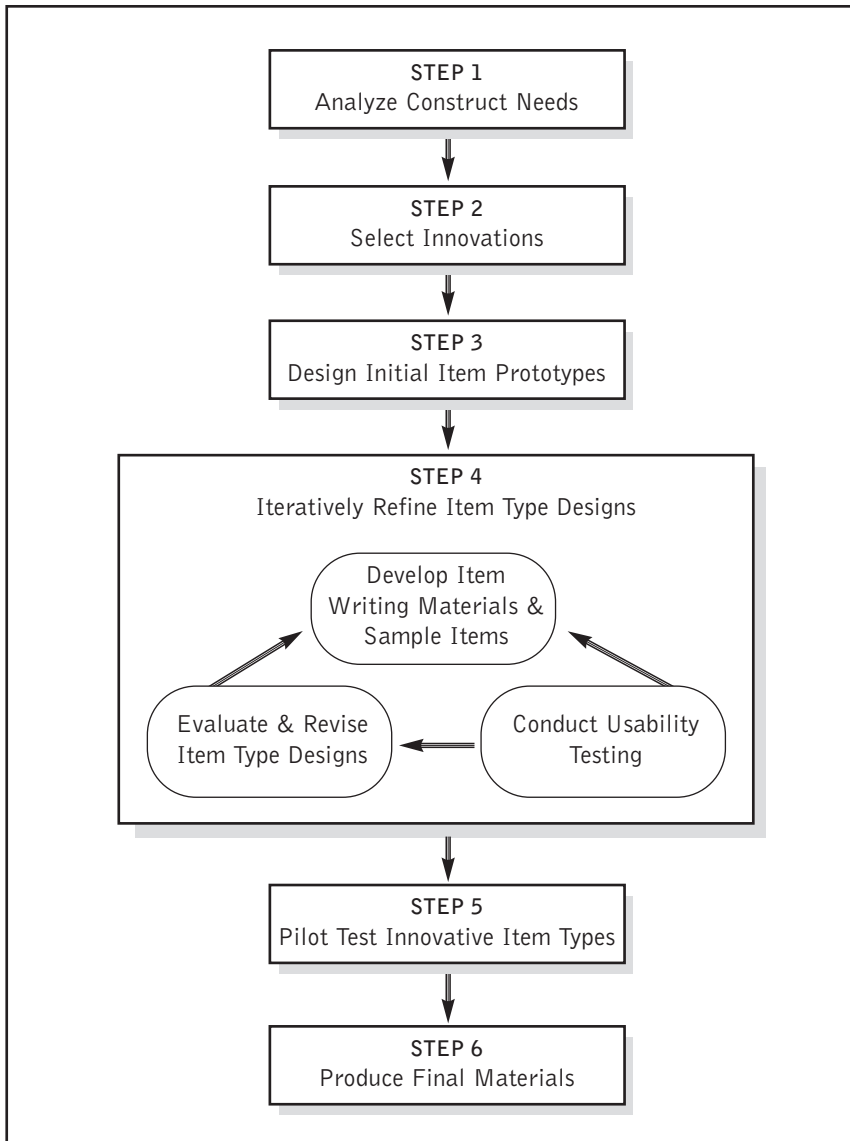


FIGURE 1. Process for Innovative Item Type Design

team needs to develop initial item writing materials and sample items for each of the approved prototypes (Step 4A). Next, the resulting sample items are used with a few members of the examinee population in usability testing of the item type designs (Step 4B). Finally, thorough stakeholder reviews that are conducted to ensure each item design can be implemented (Step 4C). Any problems that are identified, either through usability testing or stakeholder review, are jointly considered so that revisions and potential improvements can be effected. The revised item types may be returned to the content development team to begin the next round of development and review.

The test development teams can think of each proposed item type as needing to proceed iteratively through all of the Step 4 activities in successive rounds, until it has achieved a satisfactory quality criteria level. When a prototype has reached that satisfactory level, then that item type exits the feedback and revision “loop,” as it does not need any further iterations. By this point, a prototype has developed into a clearly specified item type. Some innovations may require many more iterations to reach an appropriate level of quality than others.

Step 4A – Iteratively Develop Item Writing Materials and Sample Items

The content development team will begin this step with a set of revised prototypes or item type designs. Before creating actual test items, item templates, item writing guidelines, and item writer training materials will need to be prepared for each proposed item type. The specifications that were created for an item prototype can first be translated into templates at varying levels of specificity.

In general, a template provides a framework for item construction that can serve to improve item structure, production efficiency, and exam security. Templates can help improve item structure by standardizing the way in which each new item format will be constructed and presented.

In a related fashion, production efficiency may also be improved, as item writers can be tasked with filling in components of a template, instead of creating an entirely new item concept each time. Since templates can be used as a tool for quickly developing different versions of an item, they can be useful as a security measure, depending upon the level of specificity to which the template has been designed.

A template should be created for each prototype item format. These templates can be created at various degrees of detail, and can be nested at different levels. For example,

Table 1. Steps in the design of innovative item types

<p>1. Analyze the Exam Program's Construct Needs</p>
<p>TASK: Consider the exam program and the potential value of new items types.</p> <ul style="list-style-type: none"> • Consider the exam program’s goals and requirements. • Consider those areas currently being measured well, and those that are not. • Identify any needs in the construct coverage of the current assessment. <p><i>Outcome: Identification of a few specific areas of content or cognition that ought to be added to the assessment.</i></p>
<p>2. Select Specific Innovations</p>
<p>TASK: Consider specific possible innovations and their potential match to the construct needs.</p> <ul style="list-style-type: none"> • Consider aspects of innovation: assessment structure, response action, media inclusion, interactivity, complexity, fidelity, and scoring methods. • Evaluate the match of selected innovations to the construct needs of the program. • Consider the potential costs and benefits of those innovations for the program. <p><i>Outcome: Selection of a few innovative item types for possible inclusion.</i></p>
<p>3. Design Initial Item Prototypes for Internal Discussion</p>
<p>TASK: Develop and review draft prototypes of the proposed innovative item types.</p> <ul style="list-style-type: none"> • Develop prototypes of each planned new item type, based on the construct goals and an awareness of the practical implications. • Review each prototype. <ul style="list-style-type: none"> • Evaluate whether a given prototype addresses the construct need. • Identify possible usability problems in the item interface. • Identify possible technological concerns for development or delivery. • Identify possible scoring issues or psychometric needs. • Obtain rough cost estimates for development and delivery. • Communicate concerns across all teams and determine appropriate revisions. <p><i>Outcome: Revision or elimination of each prototype.</i></p>
<p>4. Iteratively Refine the Item Type Designs</p>
<p>TASK: Refine each item type design through an iterative series of test development tasks.</p> <ul style="list-style-type: none"> • 4A – Iteratively Develop Item Writing Materials and Sample Items <ul style="list-style-type: none"> • Develop draft item templates for each item type. • Develop initial item writing guidelines for each item type. • Develop item writer training for each item type. • Have a few sample items written, making sure all planned item features are represented. • 4B – Iteratively Conduct Usability Testing <ul style="list-style-type: none"> • Conduct multiple rounds of usability testing. • 4C – Iteratively Evaluate and Revise the Item Type Designs <ul style="list-style-type: none"> • Evaluate, across all stakeholder groups, each new item type. • Determine necessary changes, based on input from usability testing and all stakeholder reviews. • Revise each item type or eliminate from further consideration. <p><i>Outcome: Approval of specific item type designs for pilot testing (perhaps after 3-5 rounds of refining).</i></p>
<p>5. Conduct a Pilot Test of the Innovative Item Types</p>
<p>TASK: Conduct a pilot test, of both item type implementation and item analysis.</p> <ul style="list-style-type: none"> • Implement a full system test of each new item type. • Collect examinee response data and conduct thorough item analyses. <p><i>Outcome: Identification of any remaining essential changes needed, or approval of an item type for operational use.</i></p>
<p>6. Produce Final Materials</p>
<p>TASK: Produce final documentation reflecting the design decisions made for each new item type.</p> <ul style="list-style-type: none"> • Produce final materials addressing: item templates, item writing guidelines, and item writer training. • Produce any other materials needed to document modifications to exam program processes and procedures. <p><i>Outcome: Full preparation of the exam program for implementation of the new item types.</i></p>

a basic template might include the structure and layout of an item format for structured tasks involving software simulations. From this general template a sub-template could be created for word processing tasks. A further sub-template might be word processing tasks involving formatting of text. Regardless of the number of template levels that are created, their use should help streamline the item writing and development process, as item writers' tasks are constrained and their responsibilities for making design decisions are reduced.

Along with item templates, new item writing guidelines should be developed for each of the proposed item designs. Standard item writing guidelines will continue to be useful, but may not be sufficient for new item types. Many innovative item types include elements that are not fully addressed in these standard item writing guidelines. Test developers and SMEs should attend to the innovative aspect of the proposed item types in order to fully specify appropriate characteristics and uses of the innovation. For example, innovative items that include graphics may produce inconsistent results if item writing guidelines fail to specify characteristics of the images and how they should be used in items. The provision of thorough item writing guidelines should strengthen the standardization and quality of the items written to each new item type design.

Once the item templates and item writing guidelines have been drafted they should be incorporated into item writer training materials. Depending upon the status of the exam program, these training materials may need to be developed for the first time. In other cases, existing content development and item writer training materials may simply need to be expanded to address the proposed new item types. An example of such an expansion might arise when item writers are asked to create questions that incorporate videos of interpersonal interactions. In this case, it may be helpful to incorporate video-based activities into the item writing session. The item writers, for whom writing video scripts is likely to be a new skill, could actually act out and videotape a few sample scripts as part of their training. A training exercise such as this might help clarify for the item writers the information and level of detail that is needed in script writing.

When the training materials have been drafted, they can be tested on a small number of SMEs. (The same SMEs can be used to help analyze the exam program's construct needs in Step 1, to initially define the item type in Step 3, and to participate in trial item writer

training in this step.) This trial item writer training activity can help the content team determine whether the templates are comprehensive, whether the item writing guidelines are clear, and whether the training activities are effective. As a final component of the training session, a few content-relevant sample items should be written for each prototype item design. These sample items should address all the basic features of each planned new item type, so that all item design elements can be considered in Steps 4B and 4C.

The evaluative information that is obtained in Steps 4B and 4C will then feed back into another iteration of Step 4, starting with Step 4A. In this follow-up round the content development team may see a need to further expand or revise the templates, item writing guidelines, or training materials associated with a given item type. After those materials have been updated, the sample items should also be revised to reflect any needed changes. Those revised sample items will be passed on once more to the following review Steps.

Step 4B – Iteratively Conduct Usability Testing

The sample items produced in Step 4A are used to conduct the first round of usability testing. The usability of a software program is the program's relative easiness to learn and to use, while usability testing involves evaluating the program to identify its usability problems (Nielsen, 2003). Usability testing is conducted widely in many software development fields, but it has particular importance for CBT applications due to its potential impact on measurement error (Harmes & Parshall, 2000). Furthermore, innovative item types within a CBT have an even greater need for usability testing because they often have more challenging item interfaces than the interfaces used in traditional multiple-choice items (Parshall, Spray, Kalohn, & Davey, 2002).

One simple but highly effective approach to usability testing is the think aloud method. In this method, the usability participant is asked to speak, or "think aloud" as he or she attempts to use the software to carry out realistic tasks. For the design of innovative item types, one round of usability testing might consist of five or six representative participants (Nielsen, 2000, 2006) who would each attempt to use the item prototypes. Working one-on-one with a participant, an observer would note all of the participant's comments and actions while he or she attempted

to respond to each aspect of the item interface.

This process is very useful for identifying areas of difficulty or confusion that users may experience with the test and item interfaces (e.g., Harmes et al., 2004; Hoffman, Harmes, & Erb, 2007; Kayser & Parshall, 2008). The think aloud method can be further used to identify aspects of the cognitive processing examinees use to respond to an item type (Wendt, Kenny, & Marks, 2007). This provides further information for improving novel item types as well as potentially contributing to validity evidence for an item type.

Whenever possible, it is highly beneficial to conduct the first round of usability testing very early in the design process. Early usability testing can inform the design decisions before extensive development has already occurred, thus reducing the number of programming changes needed. Furthermore, early identification of usability problems can enable test developers to improve each item type design quickly and cost-effectively (Bias & Mayhew, 1994).

Another important guideline for usability testing is the clear recommendation to conduct multiple rounds of usability testing (Gould, Bois, & Ukelson, 1997). That guideline is incorporated into this model for item type design in that several rounds of review and revision are planned. Each round of usability testing provides the opportunity to investigate examinees' comprehension of the new item types, as well as their ability to use the item interfaces. In the vast majority of cases, usability testing will reveal valuable refinements that can be made in the item types or their implementation.

The usability problems that are identified in each round of testing can be considered, along with other review feedback obtained in Step 4C, to identify potential improvements to the item type designs. These solutions are then implemented in the next iteration of Step 4A, and the next round of usability testing offers the opportunity to evaluate the effectiveness of each attempted solution.

Step 4C - Iteratively Evaluate and Revise Item Type Designs

In this step, the proposed item types should be evaluated by all concerned parties. All the stakeholder groups that evaluated the initial prototypes, in Step 3, should be included, as well as any other teams that might be

involved in the eventual development and implementation of the item types. A full consideration of issues might include: item banking, programming development, media development, test publishing, media delivery, test delivery, examinee seat time, scoring, and score reporting. The staff responsible for development of the CBT tutorial should also consider any new instructional or practice materials that may be needed. In many instances it can be very helpful to include a draft tutorial in the usability testing, to ensure that examinees understand any novel aspects of the item interfaces and can undertake any novel response actions required. For adaptive exam programs, additional issues might arise concerning modifications to the CAT framework that may be needed in order to include some innovative item types. Finally, the weighting of the innovative items in the final score may also need to be addressed.

After the individual teams or departments have evaluated the prototypes, a joint discussion of issues will be needed. Any problems identified, whether through these evaluations or through the usability testing in Step 4B, should be considered across all the stakeholder groups. This group discussion of issues is critical to ensure that a potential revision, intended to address problems in one area, does not create unexpected problems in another area.

This group discussion is likely to result in a set of changes to one or more of the item type designs; in some cases changes may also be needed in the item writing materials, developed in Step 4A. These revisions will feed back into another round of refining each item type design through the various activities in Step 4.

The end result of these reviews and revisions should be a set of item type designs that meet the approval of all teams. These new innovative items are then ready for pilot testing.

Step 5 – Conduct a Pilot Test of the Innovative Item Types

After three to five iterations, most innovative item type designs should be near completion. Once an item type appears to be satisfactory to all stakeholder groups, a pilot test should be conducted. The pilot testing should include a full-system test of each item type. Each of the exam program systems (e.g., item banking, test publishing, test delivery and administration, examinee response capturing, item analysis, test scoring) should be conducted for each new item type. Each team will need to confirm

that they can successfully implement all of the planned new item types.

While it is clear that this step involves considerable effort, it is also highly valuable to have every team evaluate the full implementation of each new item type. Deferring this system check until an operational “go live” date can substantially increase the cost and difficulty of addressing any problems that may occur.

Another important aspect of this step will be the item analyses conducted on the examinee response data for each item and each item type. Distractor analysis can be particularly relevant for investigating the cognitive functioning of novel item types. If all of the previous steps have been carefully carried out, it is unlikely that an item type would need to be eliminated from further consideration by this point. However, in some cases the pilot test results will reveal further item type improvements that may be worth implementing. In a few cases, these changes may be substantive enough to call for another iteration of Step 4. The end result of this step should be an approved final design for each of the new innovative item types.

Step 6 – Produce Final Materials

The goal of this last step is to finalize all of the materials needed to implement each approved item type as part of the regular exam program. All of the documentation regarding each new item type should reflect the final decisions that were made, based on the stakeholder groups’ reviews and the usability testing. Furthermore, documentation for ongoing tasks such as item writing, test development, and pool maintenance should address all of the processes and procedures needed to handle any novel elements associated with the new item types.

The result of this step should be a complete exam program, fully prepared to implement the newly designed innovative item types.

Summary

The process for innovative item type design presented here is not the only approach that can be followed to produce high quality novel item types. Nevertheless, following this six-step model should help to systematize the design and development of innovative item types, resulting in items that have high measurement quality, and are also logistically practical and acceptably affordable. The goal of

incorporating innovative items into an assessment is generally to maximize the construct benefit while minimizing the costs. However, if an exam program neglects some of the steps detailed above, the true construct benefit may not be realized and the costs may be punitively large. A thorough and thoughtful design process should allow exam programs to proceed in a way most likely to minimize costs while yielding high-quality items that truly help expand the construct measurement.

References

- Bias, R. G., & Mayhew, D. J. (Eds.). (1994). *Cost-justifying usability*. Boston: Academic Press.
- Gould, J. D., Bois, F. J., & Ukelson, J. (1997). How to design usable systems. In Helander, M., & Landauer, T.K., & Prabhu, P. (Eds.). *Handbook of human-computer interaction, 2nd, completely revised edition*. (pp. 231-254). New York: Elsevier Science Publishers.
- Harmes, J. C., & Parshall, C. G. (2000, November). *An iterative process for computerized test development: Integrating usability methods*. Paper presented at the annual meeting of the Florida Educational Research Association, Tallahassee.
- Harmes, J. C., & Parshall, C. G. (2007, February). Development and evaluation of an innovative computer-based assessment. Poster presented at the annual meeting of the Association of Test Publishers, Palm Springs, CA.
- Harmes, J. C., Parshall, C. G., Rendina-Gobioff, G., Jones, P. K., Githens, M. P., & Dennard, A. (2004, November). Integrating usability methods into the CBT development process: Case study of a technology literacy assessment. Paper presented at the annual meeting of the Florida Educational Research Association, Tampa, FL.
- Hoffman, D. J., Harmes, J. C., & Erb, J. P. (2007, April). Usability evaluation for computer-based testing software: Comparing method effects on information acquisition. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.
- Kayser, M., & Parshall, C. G. (2008, March). Building a global innovative test. Presented at the annual meeting of the Association of Test Publishers, Dallas, TX.
- Nielsen, J. (2003). *Usability 101: Introduction to usability*. Retrieved April 4, 2008 from <http://www.useit.com/alertbox/20030825.html>
- Nielsen, J. (2000). *Why you only need to test with 5 users*. Retrieved April 4, 2008 from <http://useit.com/alertbox/20000319.html>
- Nielsen, J. (2006). Quantitative studies: How many users to test. Retrieved April 4, 2008 from http://www.useit.com/alertbox/quantitative_testing.html
- Parshall, C. G., & Harmes, J. C. (2008, March). Stages in designing innovative item types. Presented at the annual meeting of ATP, Dallas, TX.
- Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. (In press). Innovative items for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice, 2nd Edition*, Norwell, MA: Kluwer Academic Publishers.
- Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer-Verlag.
- Wendt, A., Kenny, L. E., & Marks, C. (2007). Assessing critical thinking using a talk-aloud protocol. *CLEAR Exam Review*, 18(1), 18-27.

Evidence-Centered Design:

A Lens Through Which the Process of Job Analysis May Be Focused to Guide the Development of Knowledge-Based Test Content Specifications

RICHARD J. TANNENBAUM

STACY L. ROBUSTELLI

PATRICIA A. BARON

Educational Testing Service

Abstract

Little research exists on how to make the transition from job analysis to test content specifications for knowledge-based licensure tests. This paper highlights how Evidence-Centered Design (ECD) principles may be used to inform job analysis processes that are conducted to develop test content specifications. ECD principles prompt test developers and their expert committees to think about critical features, such as the purpose of the test, assumptions about the test-taking population, the intended claims to be made from test scores, and the evidence needed to support the claims. The ways in which such principles relate to the typical steps followed in a job analysis are discussed.

Key words: Job analysis, Evidence-Centered Design, test specifications, licensure

Evidence-Centered Design: A Lens Through Which The Process Of Job Analysis May Be Focused To Guide The Development Of Knowledge-Based Test Content Specifications

One purpose of licensure is to identify candidates who have the knowledge believed to be important for safe and effective practice (AERA, APA, & NCME, 1999). The *Standards for Educational and Psychological Testing* indicate that for this purpose “tests used in credentialing are intended to provide the public . . . with a dependable mechanism for identifying practitioners who have met particular standards. The standards are strict but not stringent as to unduly restrain the right of qualified individuals to offer their services to the public. Credentialing also serves to protect the profession by excluding persons who are deemed to be not qualified to do the work of the occupation” (AERA, APA, & NCME, 1999, p. 156).

A credential (license) signifies that practitioners have demonstrated the type and level of knowledge believed needed for competent professional practice. It does not mean, however, that practitioners will be competent professionals. A license is neither a guar-

antee of the public's protection nor of the practitioner's competency on the job (Schmitt, 1995). As noted by Clauser, Margolis, and Case (2007), "A passing score on a licensure examination may be seen as a *prerequisite* for acceptable practice, but not a *guarantee* of acceptable practice" (p. 717). Clauser et al. attribute this distinction, in part, to the difference between knowing and doing; that is, test takers may have the requisite knowledge, but not demonstrate that knowledge on the job. It is also the case, as Kane (2004) observes, that necessary aspects of job effectiveness may not be presented on a licensure test (e.g., character and disposition), which led him to conclude that a licensure test measures knowledge that may be necessary but not sufficient for effective practice.

Indeed, a licensure test is expected to cover only knowledge that must be present upon entry into a profession in order to practice safely and effectively (AERA, APA, & NCME, 1999). Nonetheless, it is this connection between sufficient knowledge for job performance and content knowledge covered by a licensure test that must exist in order to support the validity of licensure test scores. This link is forged through a job (or practice) analysis (Mehrens, 1995; Tannenbaum, 1999; Raymond, 2001; 2002; Raymond & Neustel, 2006). The outcome of a job analysis is often a descriptive list of categories of knowledge and specific knowledge statements that operationally define each category. The description of knowledge is developed by practitioners, and although the specific approaches to job analysis may vary, it is common for practitioners to meet as a committee to define the job-related knowledge and for a survey of the larger profession to take place to seek verification of the committee-defined knowledge domain (Knapp & Knapp, 1995). But while the reliance on job analysis to identify knowledge believed important for safe and effective entry-level practice is commonplace, how to transition that knowledge into test content specifications is less well established (Kane, 1997; Raymond, 1996).

Test content specifications articulate the knowledge areas to be measured by a test, the weight to be given to each area, and the number and types of items included on the test (Raymond, 1996). They provide direction for the developers of a test by describing the content to be measured, the form of that content and the functional requirements of the test; they are informed by the test purpose and score use, and intended test-taking population (Schmeiser & Welch, 2007). The development of test

content specifications that accurately reflect the outcomes of a job analysis play a key role in the validation of a licensure test (Kane, 1997). As noted by Raymond (2001), "If test plans are not practice related, inferences and decisions based on the test scores will not be valid" (p.390).

Previous research on the development of test content specifications has focused on approaches to determining the weights assigned to knowledge areas (Kane, 1997; Raymond, 1996; Schmeiser, 1987); linking job task descriptions to underlying knowledge requirements (Hughes & Prien, 1989); or on general approaches to constructing test specifications from the results of a job analysis (Raymond, 2001). Yet a clear consensus on how to make the transition from job analysis to test content specifications, in particular, for knowledge-based licensure tests, does not exist. What is needed is a framework that may be used to guide the transition that maintains and reinforces the connection between sufficient knowledge for job performance and content knowledge covered by the licensure test.

The purpose of this paper is to present a theoretical account of how the principles from an existing framework used to develop tests, known as Evidence-Centered Design (ECD), may be used to inform job analysis studies so that its results may be directly parlayed into the main components that comprise test content specifications. Next, therefore, is a description of the basic tenets included in ECD, followed by an explanation of how ECD-based concepts may be used to guide a job analysis study, and thus facilitate the transition to the development of knowledge-based test content specifications.

Evidence-Centered Design

"Evidence-Centered Assessment Design (ECD) is a methodology for designing assessments that underscores the central role of evidentiary reasoning in assessment design" (Mislevy, Almond, & Lukas, 2003, p. 20). The process of ECD helps to make clear the claims of the test and to weave together the claims with the observable evidence needed to support those claims, the types of items needed to elicit the evidence, and the scoring rules needed to accumulate and synthesize the evidence to draw appropriate conclusions about test takers' knowledge. As noted by Williamson, Almond, and Mislevy (2004), "This approach results in a more complete representation of the design rationale for an assessment, better targeting of the assessment for its intended purpose, and a more substan-

tial basis for a construct-representation validity argument supporting use of the assessment” (p. 14).

There are six models that form what Mislevy et al. (2003) refer to as the Conceptual Assessment Framework (CAF) of ECD: The Proficiency Model, the Evidence Model, the Task Model, the Assembly Model, the Presentation Model, and the Delivery Model. The first three models are most relevant to the development of test content specifications, and therefore will be discussed further. A full treatment of ECD is beyond the scope of this article (see, Mislevy, Steinberg, and Almond, 2003, for a complete description of ECD).

Before describing the different models, it is important to recognize that the boundaries between the models are permeable. Aspects of the Proficiency Model and the Evidence Model, for example, overlap, and need to be considered concurrently in the test development process. Similarly, some features of the Evidence and Task models overlap. The fluid nature of ECD is consistent with the fluid nature of test development, which does not necessarily occur in sequential fashion. The interaction among the models will become clearer when the discussion shifts to how ECD-based concepts may inform job analysis.

The Proficiency Model makes explicit the purpose of the test, the intended use of the test scores, and the test-taking population. It includes the delineation of the test claims and the knowledge to be measured to support the claims (Williamson et al., 2004). Knowing what to measure and why and how that information supports the objective of the test are clearly elements of a sound validity argument (Kane, 2004). The same is true for having a definition of the test-taking population, which includes explicit assumptions about their education or training, and pre-licensure experiences; these assumptions will moderate how a knowledge domain is framed and represented during a job analysis.

The Evidence Model makes explicit the kinds and levels of knowledge that must be captured by the test to enable one to conclude that the test taker has adequately demonstrated the knowledge believed important for safe and effective entry-level practice. While the Proficiency Model defines what the test taker should know, the Evidence Model defines what the test taker must show on the test to be convincing that he or she has that knowledge. Williamson et al. (2004) refer to this as the conceptual component of

the model, which feeds directly into the nature of the test items (Task Model) needed to elicit the evidence.

The Task Model makes explicit the nature and types of test items needed in order to elicit the desired evidence about what the test taker knows; it is directly informed by the [conceptual] Evidence Model. A task model is not the same as a test item; it is broader, defining a family of test items that meet certain design parameters (Mislevy et al., 2003). According to Williamson et al. (2004), a task model articulates the features of a task (what it includes and does not include), the material presented to the test takers (graphics, text, scenarios, prompts, etc.) and what the test taker is expected to produce in response to the item type. The value of the Task Model is that it reinforces the construction and inclusion of items types that are more likely to elicit evidence directly related to whether a test taker has the knowledge necessary to be a safe and effective practitioner, which is the primary objective of licensure testing. It guards against developing items, for example, that may focus too much on definitions of terms, which may play a minor role in one’s ability to practice safely and effectively.

How concepts related to these three models may serve as the conduit through which to transition from job analysis to test content specifications is the focus of the following section. We will present our vision of how this transition may occur by first describing the basic steps included in a job analysis study, and then by describing how aspects of the three models can be used to structure the study in such a way that it yields products that become the essential components included in test content specifications. Although the application of ECD to assessment design is not new (see, for example, Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002), we believe that our specific focus on the development of test content specifications will provide test developers and those involved in job analysis for knowledge-based licensure testing much needed support for and potential insights into a process that is currently not well explicated.

Job Analysis. Despite the differences in methodologies used to conduct job analysis studies, the main goal of job analysis—when conducted to support knowledge-based licensure tests—is to define a content domain of knowledge important for safe and effective, entry-level practice (AERA, APA, NCME, 1999). Therefore, a brief description of what may be considered a typical job analysis methodology is described next, followed by an explanation of how

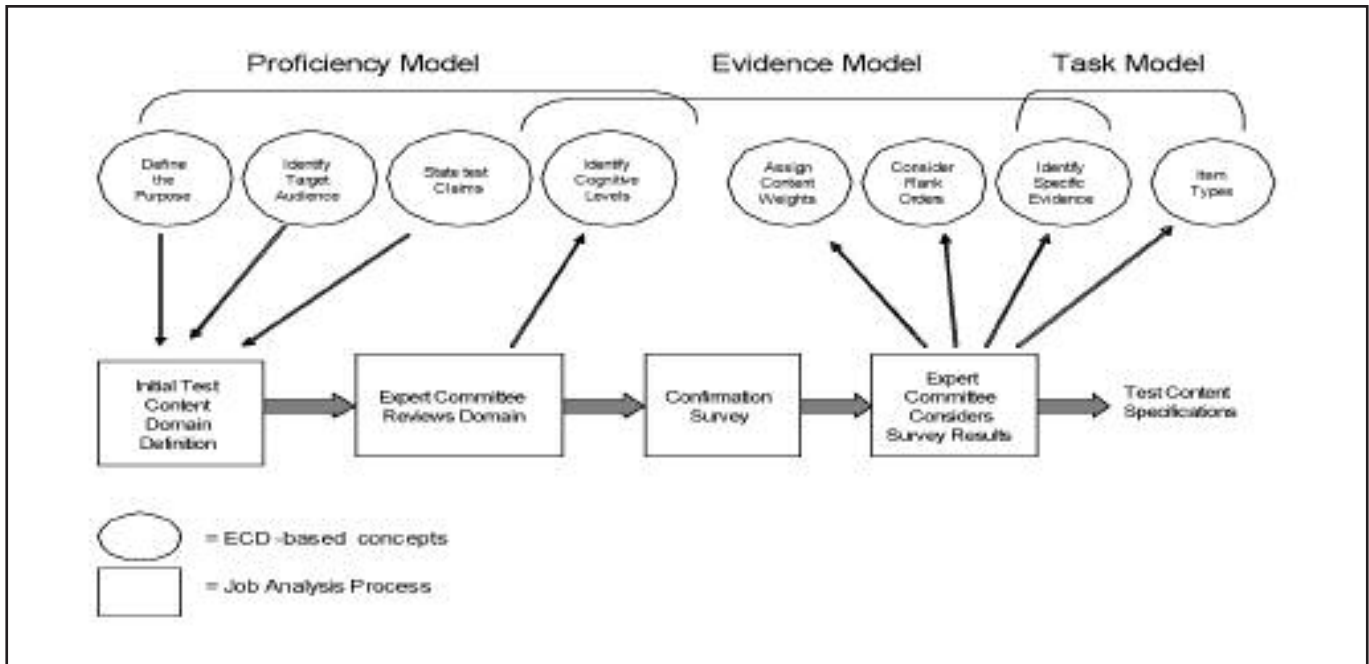


FIGURE 1. ECD-Job Analysis Framework

ECD principles (models) may be used to inform its progression.

Figure 1 contains the main steps included in our conceptualization of a job analysis study (represented by boxes), and the ways in which various aspects of ECD (represented by circles) may be used to inform each step. The brackets indicate the concepts that are drawn from each of the ECD models and where overlap occurs. The job analysis steps outlined in Figure 1 include: (1) initial test content domain definition (2) review and modification of the domain by a committee of experts (3) verification of domain importance through a survey of a national sample of experts, and (4) deliberation of the survey results by the same committee of experts, which leads to test content specifications. Next is a description of how ECD may be used to inform these steps and bridge the transition to test content specifications.

Figure 1. ECD-Job Analysis Framework

Using the Proficiency Model

The Proficiency Model may be used to inform the initial construction of the knowledge domain, which consists of a series of knowledge statements. Its value lies in making explicit three factors that must be considered to contextualize the domain properly. The model reinforces the need

to articulate the purpose of the test; to describe the target test-taking population; and to state the operational claims to be made from test scores. Essentially, these three ECD factors maximize the likelihood that construct-relevant knowledge will be included in the domain. Next is a brief description of how these components serve as a filter through which an initial test content domain may be constructed.

Defining the purpose. Domain construction for a knowledge-based licensure test is a form of domain sampling; not all knowledge associated with a field is, or should be, represented. The relevance of the knowledge statements is dependent on the purpose of the test; therefore, the first step toward developing the initial content domain is making explicit the purpose of the test. Because the purpose of a knowledge-based licensure test is to determine whether a candidate has the requisite knowledge to enter the field and to practice safely and effectively, necessary criteria for domain construction would be: Is the knowledge statement important for safe and effective practice? Is this knowledge required upon entry into the field? Knowledge more relevant to job success would be considered construct irrelevant, as the purpose of a licensure test is not to predict job success, per se. And knowledge that may be important for practice, but may be legitimately learned on the job without impeding the likelihood of

competent entering practice would also be considered construct irrelevant for licensure purposes. Working through the Proficiency Model, therefore, serves to reduce the chances that the knowledge domain strays too far from the intended purpose of the test.

Identifying the Test-Taking Population. The assumptions we make about a test taker's educational training and experience will moderate how a domain is represented and framed, and, much like defining the purpose of the assessment, is part of the validity argument for a test (AERA, APA, NCME, 1999). A test is not universally appropriate and test scores are not universally valid; validity statements need to be qualified in terms of the test purpose and the population of test takers. The assumptions made about test takers and how those assumptions interact with the purpose of the test, therefore, play a pivotal role in forming the domain. For example, one assumption that might be made about test takers for a licensure assessment for special education is that most would have earned at least a bachelor's degree in this field. By making this assumption about test takers' educational training, concerns during domain construction about test takers' exposure to and opportunity to learn "core" knowledge areas is less of a potential threat to validity.

With regard to experiential background, one assumption that might be made about test takers seeking a license as a teacher of deaf or hard of hearing students, for example, is that they do not necessarily need to be skilled in American Sign Language (ASL) in order to be considered qualified to enter the field. Therefore, including knowledge applications on the licensure test specific to ASL may be considered construct-irrelevant. What may be relevant, however, is that entering teachers know the conditions under which they must recruit an ASL interpreter to help facilitate instruction; including items on the test related to this understanding would support construct validity.

Stating Test Claims. The last component of the Proficiency Model of ECD includes identifying the claims of the test. Claims may be operational or conceptual. Operational claims are the explicit statements to be made from test scores. For example, the highest level operational claim for a licensure test is that test takers who pass the test have demonstrated a sufficient amount of job-relevant knowledge to be considered ready to enter the field or to practice without supervision. If a decision is to be made on the basis of a category score (sub-score), that category score is a lower level operational claim. If no

unique decision is intended to be tied to a category, the category is considered to be a conceptual claim. It exists to help define the overall structure of the licensure content domain, but no specific outcome or decision is based on how a test taker may have performed on knowledge specific to that category.

Clearly, there is an interaction between the test content and the cut score set by the authorized licensing agent or agency; as such, the meaning of the performance standard – what it means to be qualified to enter, which the cut score represents – becomes an embedded aspect of the operational claim. Understanding which claims are operational is essential to the construction of the initial domain. The claims are really the desired decision outcomes from the test; the only sure way then of building a test to support these desired outcomes, is to define them at the beginning stages of test development, which for licensure testing, starts with job analysis.

Overall, an understanding of the Proficiency Model provides a useful architecture for structuring and guiding the development of an initial knowledge domain. Foremost, the model requires that those involved in test development think through some key issues pertaining to validity: defining the intended purpose of the test, defining the test-taking population and assumptions about them, and stating the operational claims of the test.

Using the Evidence Model

The Evidence Model plays a prominent role at two points in the job analysis process. The first is during the stage in the process whereby an assembled committee of experts reviews and modifies the initially constructed test content domain (a point of overlap with the Proficiency Model), and the other is during the second meeting of the expert committee, which occurs after a survey of the field is conducted to confirm the importance of the modified domain (a point of overlap with the Task Model). In the context of job analysis, the Evidence Model serves to inform how the test content domain should be structured, to maintain focus on the depth of knowledge that should be expected of entering practitioners, to inform the relative weights to be applied to the different sections of the domain, and to define the behavioral indicators of evidence that need to be elicited by test items.

A knowledge domain for licensure tests is seldom unidimensional, in the sense that there is only one overarching

category of knowledge. Most often a domain consists of several sub-categories; for example, a secondary school mathematics licensure test may include sub-categories of Number Theory, Algebra, Geometry, Trigonometry, and Statistics. These categories help the expert committee to frame the knowledge domain into meaningful clusters. They may be considered conceptual claims in that they support the overall operational claim of the test, but scores on these sub-categories typically are not used to decide if a test taker has passed the licensure test.

During the expert committee's review of the initial domain, the Evidence Model also prompts the experts to be explicit in the depth of understanding they expect of entering practitioners. While it may be that Statistics is an important sub-category of knowledge (defined in the Proficiency Model) for a secondary school mathematics licensure test, the question remains, what are the appropriate cognitive levels of understanding within this sub-category? For example, is recognizing measures of central tendency sufficient, or is understanding the differences among measures of central tendency more appropriate? These reflect different cognitive demands on the entering practitioner, and have direct implications for the nature of the evidence that the test items, ultimately, will be designed to elicit. A value of the Evidence Model at this stage in the job analysis process, therefore, is that it requires the committee to engage in explicit discussions of the nature of content competency expected of the entering practitioner. These discussions, however, are also informed by the previous filters that operated during the construction of the initial domain: Is the knowledge important for safe and effective practice? Is this knowledge required upon entry into the field? These types of considerations could also have been discussed under the Proficiency Model, as they focus attention of what particular knowledge may be important to measure on the licensure test, and reflect an area of overlap between the Proficiency Model and the Evidence Model. Their discussion under the Evidence Model helps to reinforce that the expert committee should begin to consider delineating and evaluating knowledge statements keeping in mind the type of evidence they would expect to see from a test taker.

In job analysis, a survey of the practitioners often occurs after the expert committee has completed its modification of the domain. The survey is used to collect judgments of importance of the knowledge statements that form the domain for entering practice and is an opportunity to collect recommendations about the weights to be applied to

the different categories of knowledge (e.g., Number Theory, Algebra, Statistics). The outcomes of the survey are presented to the expert committee, during its second meeting, which begins the formal process of defining the specific types of evidence needed to support the operational claim of the licensure test and constructing the test content specifications.

Under the umbrella of the Evidence Model, the expert committee considers the recommended category weights from the survey and its own recommendations for the weights. The rank ordering of categories based on the weights and the absolute value of the weights are discussed in light of the operational claim of the test; that is, the discussion focuses on the most appropriate weighting scheme to support a fair and reasonable decision of sufficiency of job-relevant knowledge to enter the field. The results of the survey also enable specific knowledge statements to be rank-ordered by importance within each knowledge category. This rank-ordering provides an indication of relative importance within a knowledge category and points to particular knowledge statements that may merit greater attention during the discussions of types and sources of needed evidence.

The next objective of the expert committee is defining measurable indicators of evidence associated with the knowledge statements. The crux of the matter is defining the specific type and nature of evidence that must be elicited to provide assurance that the test taker has the knowledge believed important for safe and effective entering practice. So, for example, if the survey results confirm that "understanding differences among measures of central tendency" is important; the question now becomes, what must a test taker demonstrate on the test to attest to his or her understanding? Must he or she be able to calculate different measures of central tendency? Must he or she be able to explain when one measure of central tendency would be more appropriate than another measure? The committee needs to consider how an understanding of differences among measures of central tendency should be operationally defined on the test. Although each step in the job analysis process provides support for the validity argument of the eventual test; the ECD-based activity of defining the types and nature of evidence needed bears much of the validity burden, as the delineation of evidence provides concrete examples to item writers about what is and is not appropriate test content. Validity is supported to the extent to which the defined evidence aligns with the operational claim of the test, which may be traced through

the connection of evidence to important knowledge statements. This validity path, naturally, extends to the types of test items that are needed to elicit the required evidence; this now leads to the Task Model.

Overall, the value of the Evidence Model is that it prompts those engaged in test development to make explicit the emphasis (weight) that each knowledge area should receive on the test and also the specific type and nature of content competence that needs to be elicited from test takers to support the operational claim of the test.

Using the Task Model

The transition from the Evidence Model to the Task Model, at least for developing test content specifications, is subtle. The discussion of evidence includes the need for the evidence to be measurable. Measurable encompasses issues of reliability, fairness, and fit with the intended structure of the test. Types of evidence must be resistant to large amounts of error variance (unreliability); must be accessible to all test takers; and must be consistent with the overall design structure of the test. Then the discussion shifts to the types of items that need to be crafted to elicit the specific evidence desired. In some respects this discussion reflects a negotiation between the optimal types of items and the actual types of items that may realistically be constructed, given the overall design structure of the test. Critical to any compromise, however, is making sure that the desired evidence can be elicited by whatever item types are produced.

The Task Model prompts the expert committee to articulate the evidence that the task will elicit, the factors that may moderate the difficulty of the item, the presentation format of the item (e.g., word-problem, passage-based, figural), and the types of response information (or work products) that the item will capture. Williamson et al. (2004), for example, provide examples of a variation of a constructed-response work product; one variation is where only the test taker's answer is scored, and another is where the work shown by the test taker to arrive at the answer is part of the scoring. These item features are the building blocks for the creation of item templates from which multiple variations of an item may be generated. A template may be thought of as being an item blueprint from which parallel forms of an item are created.

The test content specifications, which are the final "deliverable" of the expert committee, reflect, therefore, the

joint influence of the Evidence Model and the Task Model. They include the specific knowledge-based evidence that is expected of the test takers and covered on the test, the emphasis (weights) placed on each of the knowledge categories represented on the test, and the types and features of items needed.

Conclusion

Little research exists on how to make the transition from job analysis to test content specifications for knowledge-based licensure tests. A notable exception is the work of Wang, Schnipke, & Witt (2005), who suggested a method for determining content outline weights (the relative distribution of items on a credentialing test) by translating task weights into knowledge weights. The purpose of this paper was to highlight how concepts from Evidence-Centered Design may function as a set of lenses through which the process of job analysis may be focused to smooth a transition to the development of test content specifications. Many of the functional aspects of ECD coincide readily with the stages of a typical job analysis and offer useful guidance to help ensure that the job analysis maintains its focus on supporting the operational claim(s) of the test. Perhaps its greatest contribution is that ECD makes explicit the criteria needed to reinforce the connections between the test content specifications and the knowledge believed to be important for safe and effective entering practice.

Validity for a knowledge-based licensure test is by and large based on a verifiable connection between the knowledge important for entry-level practice and the content domain represented on the licensure test. A job analysis provides a systematic way of defining what a job-relevant domain of knowledge may be, but does not in and of itself, unaided, lead to test content specifications. ECD offers a set of principles that may be used to guide the job analysis process. Its components prompt test developers and their expert committees to think about critical features, such as the purpose of the test, assumptions about the test-taking population, the intended claims to be made from test scores, and the evidence needed to support the claims; it is through these prompts that a bridge to knowledge-based test content specifications may be constructed. As research continues to explore how knowledge-based test specifications may be derived from job analysis, ECD would seem to be a promising area to consider.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Clauser, B.E., Margolis, M.J., & Case, S.M. (2006). Testing for licensure and certification in the professions. In R. L. Brennan (Ed.), *Educational Measurement (4th ed., pp. 701-731)*. Westport, CT: Praeger.
- Hughes, G.J., & Prien, E.P. (1989). Evaluation of task and job skills linkage judgments used to develop test specifications. *Personnel Psychology, 42*, 283-292.
- Kane, M. (1997). Model-based practice analysis and test specifications. *Applied Measurement in Education, 10*, 5-18.
- Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*, 135-170.
- Knapp, J., & Knapp, L. (1995). Practice analysis: Building the foundation for validity. In J.C. Impara (Ed.), *Licensure Testing: Purposes, procedures, and practices* (pp. 93-116). Lincoln, NE: Buros Institute of Mental Measurements.
- Mehrens, W.A. (1995). Legal and professional bases for licensure testing. In J.C. Impara (Ed.), *Licensure Testing: Purposes, procedures, and practices* (pp. 35-58). Lincoln, NE: Buros Institute of Mental Measurements.
- Mislevy, R.J., Almond, R.G., & Lukas, J.F. (2003). A brief introduction to evidence-centered design (Research Report 03-16). Princeton, NJ: Educational Testing Service.
- Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-62.
- Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15*, 363-389.
- Raymond, M.R. (1996). Establishing weights for test plans for licensure and certification examinations. *Applied Measurement in Education, 9*, 237-256.
- Raymond, M.R. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education, 14*, 369-415.
- Raymond, M.R. (2002). A practical guide to practice analysis for credentialing examinations. *Educational Measurement: Issues and Practice, 21*, 25-37.
- Raymond, M.R., & Neustel, S. Determining the content of credentialing examinations. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of Test Development* (pp. 181-224). Mahwah, NJ: Lawrence Erlbaum.
- Schmeiser, C.B. (1987, April). *Effects of translating test analysis data into test specifications*. Paper presented at the National Council on Measurement in Education, Washington, DC.
- Schmeiser, C.B., & Welch, C.J. (2007). Test development. In R. L. Brennan (Ed.), *Educational Measurement (4th ed., pp. 307-353)*. Westport, CT: Praeger.
- Schmitt, K. (1995). What is licensure? In J.C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 3-32). Lincoln, NE: Buros Institute of Mental Measurements.
- Tannenbaum, R.J. (1999). Laying the groundwork for a licensure assessment. *Journal of Personnel Evaluation in Education, 13*, 225-244.
- Wang, N., Schnipke, D., & Witt, E.A. (2005). Use of knowledge, skill, and ability statements in developing licensure and certification examinations. *Educational Measurement: Issues and Practice, 24*, 15-22.
- Williamson, D.M., Almond, R.G., & Mislevy, R.J. (2004). *Evidence-centered design for certification and licensure*. CLEAR Exam Review, Summer, 14-18.

CLEAR

403 Marquis Avenue
Suite 200
Lexington, KY 40502

NON PROFIT ORG.
U.S. POSTAGE
PAID
Lexington, KY
Permit No. 1