CLEAR Exam Review

Volume XXVII, Number 2
Winter 2017-18

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

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

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

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

CLEAR Exam Review

VOLUME XXVII, NUMBER 2 WINTER 2017-18

Contents

 ISSN 1076-8025

Abstracts and Updates

GEORGE T. GRAY, Ed.D.
testing and measurement consultant

Rather than focusing on "breaking news" in this issue's column, we look at a resource that covers a span of thirty years to the present. The National Council on Measurement in Education's "Instructional Topics in Educational Measurement Series" (ITEMS) began as a feature in *Educational Measurement: Issues and Practice* in 1987 (NCME). These instructional modules have been authored by leaders in the measurement field. To date 45 modules have been published covering a broad range of topics, and they are accessible to the public on the NCME's website. Just Google **NCME ITEMS** and click on the web page link. All of the modules are available as downloadable PDFs.[1]

For the most part, each specialized topic is covered once in the series of 45 modules and not repeated unless there is a psychometric update. The more basic modules, such as introductions to item response theory and classical equating methodology, were covered many years ago, but the content is still current. Recent publications have been mentioned in past columns. The purpose here is to give attention to some of the "oldies but goodies" that are relevant to licensure and certification testing. In some cases, references to software or technical capabilities may be dated, but the conceptual part of the content remains solid.

Obviously not everyone is involved in crunching numbers, but the licensure and certification community as a whole benefits from shared understanding of concepts and precise use of technical language. The modules that are mentioned here are by no means the only ones that are recommended, but the focus is on older publications, basic topics, and some takeaway points that might be immediately useful. The ITEMS papers are focused on teaching and learning, and as a bonus they include a set of self-assessment questions at the end. For the purpose of this review, no attempt has been made to capture the full scope of the papers. Preference has been given to the points that are likely to be relevant to readers of *CLEAR Exam Review*.

Understanding Reliability

Ross E. Traub and Glenn L. Rowley
Module 8, Spring, 1991

Reliability is described as "the relative consistency of test scores" (p. 37).[2] This paper begins with a kind and gentle introduction to reliability, pointing out usage in everyday life before discussing reliability of test scores. Then plots of scores on two test forms are presented to illustrate the degree of correspondence between the two. The concept of an observed test score as having two components, a true score and an error component, is introduced. In a practical application of calculation of a reliability index,

1 As of April 2018, the NCME has launched an educational portal centered around the ITEMS modules. Access the portal and sign up for a free account at https://ncme.elevate.commpartners.com.

2 Page numbers refer to the numbering in the original issue and are provided for convenience in locating precise quotations in the modules as presented online.

the Kuder-Richardson formula 20 is introduced for test items scored as right or wrong (1 or 0). Factors affecting reliability are discussed, and the authors remind the reader that the question, "What makes a test reliable?" is "actually the wrong question, since a test by itself is neither reliable nor unreliable. When a test is used to assign scores to individuals, the scores that are obtained may be reliable or unreliable; it is the scores that have the property of reliability and not the test itself" (p. 42).

Standard Error of Measurement

Leo M. Harvill
Module 9, Summer, 1991

The abstract states, "The standard error of measurement (SEM) is the standard deviation of errors of measurement that are associated with test scores from a particular group of examinees. When used to calculate confidence bands around obtained test scores, it can be helpful in expressing the unreliability of individual test scores in an understandable way. . . . Interpreters should be wary of over-interpretation when using approximations for correctly calculated score bands. It is recommended that SEMs at various score levels be used in calculating score bands rather than a single SEM value" (p. 33).

Although the concept of different standard errors of measurement at different score levels is a critical (and underappreciated) concept, Harvill provides a number of illustrations to facilitate the understanding of the concept of SEM and its relationship to the standard deviation and reliability of scores. The SEM is equal to the standard deviation times the square root of the quantity *1* minus reliability. Thus, if reliability is *0*, the square root of *1* minus *0* equals *1*, and the SEM = the standard deviation times *1*. That is, the SEM equals the standard deviation. In the hypothetical opposite case, the reliability is perfect (*1*), and the square root of the quantity *1* minus *1* (*0*) is multiplied by the standard deviation. So reliability of *1* means an SEM of *0*.

The SEM can be estimated from the square root of test length. The estimated SEM for a test of 100 items would be 10 items. Using the same formula, 7 would be the estimated SEM for a 49-item test, and 5 would be the SEM for a 25-item test. Note that shorter tests have a proportionately higher SEM of scores compared to test length.

Another key concept is the precision of individual scores. An interval of plus or minus an SEM around the observed score captures the individual's true score 68% of the time. An interval of plus or minus two SEMs captures the true score 95% of the time. There are two major implications of this relationship. First is that the individual's true score frequently lies in a fairly large range. Second, there is usually no basis for inferring that the true scores for two individuals whose observed scores are a point or two apart are in fact different.

Traditional Equating Methodology

Michael J. Kolen
Module 6, Winter, 1988

For non-psychometricians, equating of test forms is frequently not well understood. This article is intended to provide an understanding of three methods of equating using assumptions of classical test theory: mean equating, linear equating, and equipercentile equating. Kolen describes equating as "an aspect of a more general scaling/equating process. In this process, a scale for reporting scores is established at the beginning of a testing program (or at the time that a test is revised). This score scale is chosen to enhance the interpretability of scores by incorporating useful information into the score scale so as to avoid misleading interpretations. . . . Score scales typically are established using a single test form. For subsequent test forms, the scale is maintained through an equating process that places scores from subsequent forms on the score scale that was established initially" (p. 30).

One of the most common forms of equating for certification and licensure examinations that use classical test theory is the Common Item Nonequivalent Groups design. A simple use of this model is for a new test form with a new group of candidates. A bridge of common items similar in content distribution to the total test (typically 20% of the length of the test form) is selected from a previous form. These items appear on the new test. Results on the common items determine whether the new candidates are of similar ability or of slightly higher or lower ability. Once this has been established, the non-common items determine whether the new test form is of the same or a higher or lower level of difficulty. A higher difficulty form will entail a lower raw passing score and vice versa. For both test forms the scaled passing score will be the same. This model is also applicable when multiple test forms are released at the same time. Note that administration data on the most recent form must be obtained before scores can be issued.

In addition to providing an introduction to equating, this paper also explains common score conversions used in equating: mean equating, linear equating, and equipercentile equating. Mean equating adjusts for differences in group means. Linear equating adjusts for differences in mean and standard deviation. And, Kolen states, "Equipercentile equating provides for even greater

similarity between distributions of equated scores than does linear equating. In equipercentile equating, scores on Form 1 and Form 2 with the same percentile rank for a particular group of examinees are considered to indicate the same level of performance" (p. 33). Formulas and illustrations complement the text.

The article closes with a discussion of equating error: both random and systematic equating error. Awareness of measurement error is a key aspect of psychometrics.

Comparison of 1-, 2-, and 3-Parameter IRT Models

Deborah Harris
Module 7, Spring, 1989

For anyone who wants to know the answer to the question, "What's this IRT about, anyway?" this is an excellent introduction. The abstract states, "This module discusses the 1-, 2-, and 3-parameter logistic item response theory models. Mathematical formulas are given for each model, and comparisons among the three models are made. Figures are included to illustrate the effects of changing the a, b, or c parameter, and a single data set is used to illustrate the effects of estimating parameter values (as opposed to the true parameter values) and to compare parameter estimates achieved through applying the different models. The estimation procedure itself is discussed briefly. Discussion of model assumptions such as dimensionality and local independence can be found in many of the annotated references" (p. 35).

The illustrations of the three different IRT models from the article (p. 36) are provided above. They vary by the number of parameters included in the model (*a; a & b; or a, b, & c*). For the 1-parameter model, the shape of the graph for all items is identical, but they appear at different points on the x-axis of the graph depending on level of difficulty of the item (designated as the *a* parameter). All items that fit the model are treated as having the same level of statistical discrimination (the *b* parameter). The curves for the 2-parameter model typically have different slopes, reflecting differences in item discrimination. The distinguishing feature of the 3-parameter graph is that the probability of correct response in the lower left corner of the graph never goes down to zero. The lowest point to the y axis is the c parameter.

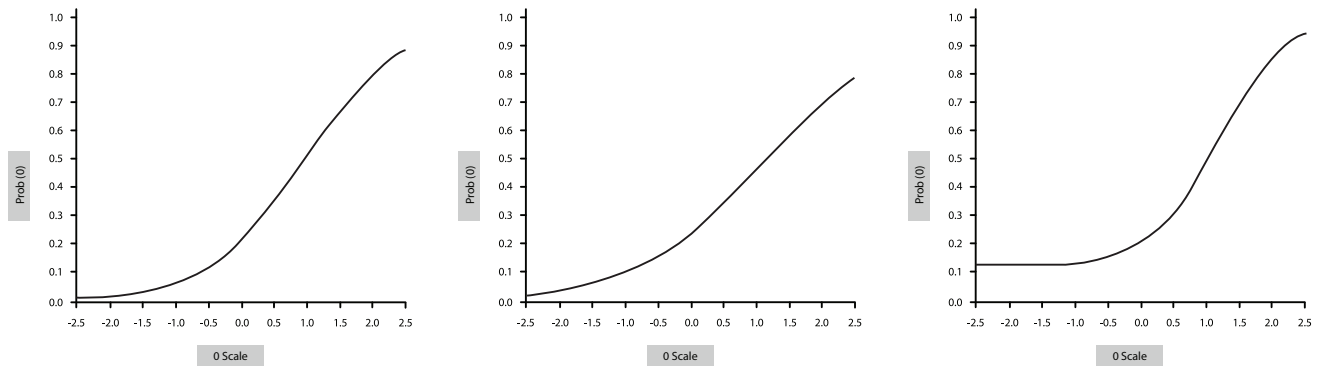
IRT Equating Methods

Linda L. Cook and Daniel R. Eignor
Module 10, Fall, 1991

The authors remind the reader that no equating method will be able to equate different tests, only multiple forms of the same test that are similar in level of difficulty, reliability, and test content.

Once an IRT model has been chosen (1-, 2-, or 3-parameter), calibrations are done to put the items on the IRT scale. For certification and licensure testing, this usually means calibration of a single test form as a first step. If there are multiple forms of the same test that have common item links, they can also be calibrated. Following the logic of this process, a bank of items that is produced by adding new forms through classical equating can be moved to IRT relatively quickly. If the test forms have not been equated, the IRT calibrated bank must be developed one test form at a time through linked items with a previously calibrated form. Once a bank of items has been equated, *pre-equated* forms can be assembled, as all scored

FIGURE 1. Examples of item characteristic curves derived from the 1-, 2-, and 3- parameter models



items on new forms will have statistics. New items can be pretested and calibrated once sufficient candidate volume is available.

The authors describe a three-step process for moving to IRT: selecting a design, placing parameter estimates on a common scale, and equating test scores. Formulas are provided with the text describing these steps, along with accompanying illustrations.

Cook and Eignor offer four practical advantages of IRT equating over classical equating: (1) "IRT equating offers better equating than that offered by classical methods at the upper ends of score scales where important decisions are often made;" (2) "IRT equating affords greater flexibility in choosing previous forms of a test for equating purposes. . . . [I]t is possible to equate a new test form (once its parameter estimates have been placed on the same scale) to any or all of the old test forms;" (3) "Re-equating is easier should it be decided to not score an item after the test is administered;" and (4) "IRT equating offers the possibility of item level pre-equating, or deriving the relationship between the test forms before they are administered operationally" (p. 42).

Comparison of Classical Test Theory and Item Response Theory and their Applications to Test Development

Ronald K. Hambleton and Russell W. Jones
Module 16, Fall, 1993

The beginning of this article explains the differences between test theories and models. Following this section, classical test theory is defined as "a theory about test scores that introduces three concepts—test score (often called the observed score), true score, and error score" (p. 40). The frequently cited relationship of observed score equals true score plus error is mentioned, highlighting the fact that the equation has two unknowns and "is not solvable unless some simplifying assumptions are made. The assumptions in the classical test model are that (a) true scores and error scores are uncorrelated, (b) the average error score in the population of examinees is zero, and (c) error scores on parallel tests are uncorrelated. In this formulation, where error scores are defined, true score is the difference between observed score and error score" (p. 40).

The authors note that "most of the work in classical test theory has focused on models at the test-score level (in contrast to item response theory). That is, the models have linked test scores to true scores rather than item scores to true scores" (p. 40). They then refer to classical item statistics (difficulty and statistical discrimination) but point out that ". . . one main shortcoming is that they are sample dependent, and this dependency reduces their utility" (p. 40).

The summary statement of this section is as follows: "Advantages of many classical test models are that they are based on relatively weak assumptions (i.e., they are easy to meet in real test data) . . . On the other hand, both person parameters (i.e., true scores) and item parameters (i.e., item difficulty and item discrimination) are dependent on the test and the examinee sample, respectively, and these dependencies can limit the utility of the person and item statistics in practical test development work and complicate any analyses" (p. 40).

The section on IRT contains an introduction to this topic, including illustrations of item characteristic curves and item information functions. In comparison with classical test theory, assumptions are strong, and IRT can be used if the model fits the data. The IRT theta scale representing both candidate ability and item difficulty level conceptually goes from minus infinity to plus infinity, but in practice, the ranges are usually single digits above and below the zero midpoint. In the authors' opinions, minimum sample size for effective use of classical test statistics is about 200, but IRT requires about 500.

In summary, IRT offers advantages, provided model assumptions are met and the candidate sample size is adequate.

Using Statistical Procedures to Identify Differentially Functioning Test Items

Brian E. Clauser and Kathleen M. Mazor
Module 19, Spring, 1998

The authors remind the reader that "differential item functioning is present when examinees from different groups have differing probabilities or likelihoods of success on an item, after they have been matched on the ability of interest" (p. 281).

They state, "Test results are routinely used as the basis for decisions regarding placement, advancement, and licensure. These decisions have important personal, social, and political ramifications. It is crucial that the tests used for these decisions allow for valid interpretations. One potential threat to validity is item bias. When a test item unfairly favors one group over another, it can be said to be biased. Such items exhibit *differential item functioning* (DIF), a necessary but not a sufficient condition for item bias" (p. 281).

Matching groups on ability for an analysis is not as easy as it might seem, for it is not simply a matter of matching on total test scores. As an example, the authors cite a test of mathematics word problems where obtained scores on the test are a mix of mathematical ability and reading comprehension. If a second ability is associated to the ability

of primary interest, the question is "whether that second ability is relevant to the purpose of testing" (p. 281).

Clauser and Mazor also make it clear that potentially offensive content, e.g., gender stereotyping, is an important but separate issue from DIF. Items should be reviewed for both DIF and fairness.

A number of examples are provided with graphs comparing IRT item characteristic curves of focal and reference groups; however, "(t)he limitation of IRT methods is that the data must meet the strong (unidimensionality) assumption of the models. These methods also require large examinee samples for accurate parameter estimation if the two- or three-parameter model is used" (p. 284).

In addition to IRT methods, several other approaches to identifying DIF are discussed: the Mantel-Haenszel statistic, the standardized difference of the proportion correct (standardization procedure), the SIBTEST computer program, and logistic regression. DIF analysis for items with polytomous scoring is also covered.

Standard Setting I: Traditional Methods

Gregory J. Cizek
Module 18, Summer, 1996

Cizek refers to setting a standard of performance as implementing "a process that identifies a point on a score scale that divides the observed test score distribution, resulting in classifications such as master/nonmaster, pass/fail, or certify/deny certification" (p. 20). Methods discussed in the module include test-centered methods, examinee-centered methods, and compromise methods.

The test-centered methods discussed vary but typically involve a panel of judges making determinations of which test items minimally proficient examinees would answer correctly or what percentage of minimally proficient candidates would answer correctly. There is considerable subjectivity in this process, but for the most part, it is superior to other methods such as setting a minimal percent correct independent of the particular test items or basing the passing score strictly on a distribution of obtained scores.

Descriptions are provided of four of the most popular test-centered methods from a historical perspective: Nedelsky (1954), Ebel (1971), Angoff (1971), and Jaeger (1982). The Nedelsky method involves judgments of which multiple-choice item options a minimally competent candidate would answer correctly. Ratings yield a decimal value for each item. These values averaged over items and raters provide a percent-correct passing score. The Angoff method is actually two methods, one mentioned in the text of the cited book chapter and the other cited in a footnote. In both cases, panelists judge whether minimally proficient candidates will answer an item correctly or not. In the most popular method (described in the footnote), the judgment is based on the percentage of minimally competent candidates who would be expected to answer the item correctly. The other method assigns a 1 or 0 to each item, depending on whether a minimally competent candidate would be expected to answer correctly or not. In either case, results are averaged over items and raters to obtain a passing standard.

Cizek illustrates the Ebel method with a table of sample data. Items are judged as "essential, important, acceptable, or questionable" and classified as "easy, medium, or hard." For each of these twelve categories (4x3), a percent correct required for mastery is set. The number of items in each category is then determined and through multiplication, addition, and finally division by the number of items on the test, the passing score is determined.

The Jaeger method asks judges to identify which items *every* competent candidate should get correct. Different panels of stakeholders complete this task, and the median value across panels is used as the passing score.

Examinee-centered methods and compromise methods are also presented. Examinee-centered methods focus on judgments about proficiency of candidates rather than performance of candidates on test items. Two methods are described: the Contrasting Groups Method and the Borderline Group Method. The Contrasting Groups Method requires the identification of two groups of individuals, one described as masters (competent) and another group not likely to meet standards for content knowledge (non-masters). Both groups take a test, and the distributions of scores are plotted. The point at which there is minimal overlap of the distributions is chosen as the passing score for the examination. While this mathematical decision procedure is appealing, the judgment criterion for selecting or setting up the groups of masters and non-masters may be challenging.

Cizek's description of the Borderline Group Method reads as follows: "Zieky and Livingston (1977) proposed using a single group judged to be at the borderline separating competent from non-competent performance. To implement the procedure, participants who are familiar both with examinees at this level and with the knowledge or skills to be tested identify a sample of members at this subpopulation. The median score of this sample can be used as a recommended standard" (p. 25).

Two compromise methods are featured. As Cizek indicates, for Beuk's (1984) method, "each participant in the standard-setting procedure is asked to make two judgments: (a) the minimum level of knowledge required to pass an

examination, expressed as a percentage of total raw score on the test, and (b) the passing rate expected, expressed as a percentage of the total population. When the examination has been administered, these expectations can be compared with reality" (p. 26). The Hofstee method involves asking judges to provide a minimum and maximum percent correct for passing and a minimum and maximum acceptable failure rate, also expressed as a percentage. These numbers are used as coordinates to define a straight line on a graph. The cumulative frequency distribution of scores is also plotted on the graph. The place where the frequency distribution and the line representing judgments intersect is the recommended passing score.

Standard Setting II: Setting Performance Standards: Contemporary Methods

Gregory J. Cizek, Michael B. Bunch, and Heather Koons
Module 22, Winter, 2004

This module examines standard setting in light of the 1999 version of the *Standards for Educational and Psychological Tests* (AERA/APA/NCME, 1999) and updates since the earlier module was published. The authors note that "often-hypothetical conceptualizations" of competence "remain important" regardless of whether a method is examinee-centered or test-centered (p. 35). As far as specific methods are concerned, the Bookmark method is described in detail. Briefly summarized, the Bookmark method involves placing the items on a test form in an "ordered item booklet" based on item difficulty. Items are calibrated using item response theory. The judges consider the likelihood of a correct response on the items by a minimally competent candidate. Immediately after the page for which a 67% probability of a correct response is estimated, the judge places a bookmark. This represents the judge's opinion of the cut score. Different judges' opinions are reconciled to obtain an approved passing score.

An Extended Angoff method is covered as well as the Angoff "Yes/No" method. The latter has an unusual history as it was the method described in the text of Angoff's 1971 chapter but received little attention. The method that captured attention and gained great popularity was offered in a short footnote on the same page. Several holistic standard setting methods are also described by Cizek et al., including the Body of Work method.

Other Recommended Modules

The emphasis here has been to highlight some basic measurement topics supported by modules published a number of years ago. A number of recently published modules have been mentioned previously in this column. That having been said, attention should be given to three modules associated with subscores and published between 2011 and 2014 (Modules 32, 37, and 38). Tong and Kolen's overview of scaling procedures (Module 31, 2010) would also be high on the priority list.

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.

Angoff, W.H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike, (Ed.), *Educational Measurement* (pp. 508-600). Washington, DC: American Council on Education.

Beuk, C.H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement, 21*, 147-152.

Ebel, R.L. (1972). *Essentials of Educational Measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Jaeger, R.M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4*, 461-475.

National Council on Measurement in Education (NCME). *Educational Measurement: Issues and Practice*. Hoboken, NJ: Wiley-Blackwell.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*, 3-19.

Zieky, M.J. and Livingston, S.A. (1977). *Manual for setting standards on the Basic Skills Assessment Tests*. Princeton, NJ: Educational Testing Service.