CLEAR
Exam Review

Volume XXIII, Number 2
Fall 2012

A Journal

CLEAR Exam Review

VOLUME XXIII, NUMBER 2 FALL 2012

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 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 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.

Contents

COLUMNS

ARTICLES

What's in a Score? Principles and Properties of Scoring

JERRY L. GORHAM, PhD
Dr. Gorham is Senior Research Scientist at Pearson VUE

ADA WOO, PhD
Dr. Woo is Associate Director of Measurement and Testing at National Council of State Boards of Nursing

KAREN SUTHERLAND, PhD, RN
Dr. Sutherland is Lead Content Developer at Pearson VUE

Most of us grow up in the company of test scores. "How did you do on your spelling test? I had an 84%." But, until one takes a course on measurement theory or works in the testing industry, one may not give much formal thought to scores as measurement theorists consider them. Scoring is in fact a very important and complex responsibility of testing professionals and deserves some attention. In this short article, we will try to make some of these scoring concepts explicit and to delineate some of the differences among types of scores and properties of scores and scales.

Defining scores

At its most basic level, a score is simply an assignment of a relative value to an examinee's response or set of responses. Usually a score is assigned to each item and then combined over the set of items that the examinee has answered. The assignment is often numerical, but does not necessarily have to be numerical; it can be a category score that represents some qualitative scale (i.e., poor, sufficient, excellent). Scores also need to reflect a quantity of knowledge on some underlying or latent scale. Scores such as these are most often derived from applications of item response theory to examinee response data. Higher scores are usually interpreted as representing "more" of the ability or construct, but not necessarily the same unit amount. For instance, the difference between a 30 and 50 may not represent the same amount of knowledge between a 50 and 70. What matters is that 30 is less than 50 which is less than 70, so that a score of 70 represents more knowledge, skill or ability along the continuum of the construct being measured. Recall the traditional distinctions usually made in introductory measurement courses between the types of scales: nominal, ordinal, interval, and ratio (Stevens, 1946). Although these distinctions may be somewhat simplifying, they are useful in understanding the types of claims that can be made from each scale type, and the types of analytical functions that can be derived legitimately from each scale type. The two scales types most useful in standardized tests are the ordinal and interval scales, each of which implies order-preservation under transformations. This property becomes very useful in reporting scaled scores and providing appropriate interpretations for test users.

Notice that in this process of assigning a relative value to a response, the test designer (usually a content expert) is making a judgment about the examinee's response in relation to the construct being measured. This judgment is an expert interpretation of how much knowledge, skill, or ability the examinee has demonstrated; not all experts may agree with that judgment, therefore committees of experts often serve as review panels to ensure some consensus among representative experts about the scoring judgment being applied.

Types of scores

In standardized testing, the most common item score type traditionally has been the dichotomous item score. A dichotomous score is one in which there are only two judgments made about an examinee's response: either the score is right or wrong. As a result, usually the values assigned are: full credit or no credit, and the numerical values of 0 or 1 are commonly used for scoring at the item level. At the test level, these are usually called "pass" or "fail" scores because the examinee has demonstrated sufficient knowledge, ability or skill on the examination to receive full credit or no credit for the specific purposes of the test.

Polytomous scores refer to "many cuts" (poly + temnein, Greek) and are commonly called partial credit scores in classroom tests. If an item is scored in more than two categories, it can be considered a polytomous score. These categories can range from three to many, and are often chosen based on what is reasonable for a rater or judge to adequately distinguish, as well as on the level of desired rater consistency across all candidates in assigning the predefined rubrics. One principle behind rubric development is that differences in score levels assigned need to be recognizable and explicit. If experts cannot clearly define the characteristics of a response that distinguish its score category from the category above or below it, then the rubric development process is unclear and needs refinement. Examinee essays are an example of items that are often graded polytomously. These items may require longer examinee effort, response time, and generally produce more statistical information than discrete multiple choice items, so may be given a higher weight overall or are even reported separately because of the effort required and inherent measurement value. In addition, items that are scored polytomously often require more effort on the part of content experts to develop the rubrics for scoring.

Properties of good scores

One important property of scores is the ordering property that represents logical increments of ability or skill along some continuum that is understandable to both experts and non-technical users. Although a dichotomous scoring rubric may seem somewhat harsh by resolving all examinee responses as either "correct" or "incorrect", suggesting that an examinee either knows all or nothing about a stimulus, this method is only severe at the item level; when applied to an entire group of items on an examination, the combined total score represents an ordering of values that should be roughly equivalent to an ordering of knowledge along the construct that is being tested. If the data being modeled are making use of item response theory, then that ordering will usually be even more precisely defined than for raw score totals or converted score scales (Nunnally & Bernstein, 1994).

A second property of a good scoring system is that scores need to be combine-able into scales for the entire exam. Sampling many items in a construct usually lends itself to some numerical combination and transformation into a scale for that construct. But, if scores are measuring very different types of constructs, it does not make much logical sense to combine those scores into an entire composite score even if the mathematics permits it (Thissen & Wainer, 2001). Rather, if reading comprehension and mathematical skills are being tested, those scores are better kept separate so that test users can interpret their meaning clearly.

A third property of scores is that scores need to be comparable from one time event to another. In other words, a score of X on an exam should represent approximately the same level of ability on a construct regardless of which version of the exam the test taker received or when the examinee took an exam. In order to conceptually evaluate comparability, measurement experts think of items, subtests, and entire tests as having properties of "difficulty". If one version of an exam is more difficult than another version, no reasonable person would expect examinees of equivalent ability to receive the same scores on each version of the exam. These innate differences in difficulty on exam versions lead to the broad area of measurement procedures known as linking, equating, and model calibration to adjust for these differences in difficulty and to ensure that a score level can be interpreted consistently across time and across different test versions (Kolen & Brennan, 2004). Testing professionals spend an enormous amount of time and effort to ensure that differences in test versions do not lead to differences in reported scores for equivalent-ability examinees.

A final important property of scores is ease of use. Scores are not helpful to consumers of test products unless they can be interpreted easily and clearly by the multiple users of tests (examinees, constituents, educators, regulators and policy makers, measurement professionals and researchers). "What does my score mean and how is it related to other scores on the scale?" is a basic question that should be addressed when reporting scores. Reported scores often use qualitative explanations for levels of achievement or make use of norms, percentile ranks, or some other measures of comparison that will assist users in understanding and using scores appropriately.

Scoring in the 21st century

In the past few decades, scoring has evolved from what was most often a simple, straightforward use of dichotomous scoring for multiple-choice items to the more complex polytomous scoring for multiple-response items, and rater-based scoring of short answer and extended answer item types. In addition to multiple-choice items, case studies and testlets are some of the likely item types for high-stakes standardized tests of the future. These alternate formats provide insight into an examinee's thought processes, reasoning, logic and composition in ways that short answer or multiple-choice items could not. Although these item types may be more expensive to administer and more complex to score, both human raters and automated rater scores have demonstrated some success at providing consistent scores that are more efficient and cost effective to produce.

Innovation in item types tends towards authenticity of the sample being collected and increasing complexity of the sorts of tasks required by the exam. The use of models and methods to accommodate highly related items and dynamic branching based on examinee responses may be one path for future testing. Automated scoring systems for essay items are becoming increasingly more practical and cost effective (Williamson, Mislevy, & Bejar, 2006). Complexity of related tasks and dynamic routing, processing and scoring of those tasks would appear to be the future for testing. Although the development of Bayesian networks and neural net models are primarily in the theoretical stages at this time, it is expected that applications of these models will become more prevalent in the near future and will yield realistic and authentic features to standardized tests and performance tasks that are not currently available.

References

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices.* New York, NY: Springer.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory.* New York, NY: McGraw-Hill.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103* (2684), 677-680.

Thissen, D., & Wainer, H. (2001). *Test Scoring.* Mahwah, NJ: Erlbaum.

Williamson, D., Mislevy, R., & Bejar, I. (2006). *Automated Scoring of Complex Tasks in Computer-Based Testing.* Mahwah, NJ: Erlbaum.