Resources for Assessment Development

A Bibliography for the Assessment Community

Prepared for the National Council on Measurement in Education

January, 2010

Ian Hembry and Anthony Fina

University of Iowa

Resources for Assessment Development
A Bibliography for the Assessment Community

Resources for Assessment Development
A Bibliography for the Assessment Community

Introduction

In 2010, we are in a period of change in how we conduct educational testing in the United States. The increased availability of technology in our schools, the integration of cognitive science into test design, and the ongoing development of better methods for assessing the broad, diverse groups of students in our country, along with the US Department of Education's focus on innovation in assessment has resulted in expanding our thinking about how we develop large-scale assessments. This annotated bibliography provides resources related to current methods of test development; it is not meant to be exhaustive but instead representative of a wide variety of resources. We expect that the bibliography will grow steadily over the next few years.

Test Design references cover multiple aspects of structuring a testing program, from the theoretical base through the intended uses.

Item Development references provide guidance for developing, reviewing, and evaluating multiple-choice and constructed-response items, including item rubrics.

Test Assembly and Evaluation resources focus on determining the quality of test items and the test as a whole.

Computer Related Topics address the comparability of computer-delivered tests to paper and pencil test and computer adaptive testing.

Scoring addresses holistic and analytic scoring process for constructed-response items.

General resources are widely-accepted testing industry standards and guidelines.

Testing Program Documents are resources from national testing programs that contain information about test development.

1. Test Design

General Topics and Concerns

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

There are many threats to validity in high-stakes achievement testing. One major threat is construct-irrelevant variance (CIV). This article defines CIV in the context of the contemporary, unitary view of validity and presents logical arguments, hypotheses, and documentation for a variety of CIV sources that commonly threaten interpretations of test scores. A more thorough study of CIV is recommended.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models (Module 7). *Instructional Topics in Educational Measurement* (ITEMS). Madison, WI: NCME. Retrieved March 01, 2009, from http://www.ncme.org/pubs/items/13.pdf

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 s 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 though 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 (e.g., Hambleton, 1988).

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 3*(4), 379-416.

Educational test theory consists of statistical and methodological tools to support inference about examinees' knowledge, skills, and accomplishments. Its evolution has been shaped by the nature of users' inferences, which have been framed almost exclusively in terms of trait and behavioral psychology, and focused on students' tendency to act in prespecified ways in prespecified domains of tasks. Progress in the methodology of test theory enabled users to extend the range of inference and ground interpretations more solidly within these psychological paradigms. Developments in cognitive and developmental psychology have broadened the range of inferences we wish

to make about students' learning to encompass conjectures about the nature and acquisition of their knowledge. The same underlying principles of inference that led to standard test theory can support inference in this broader universe of discourse. Familiar models and methods--sometimes extended, sometimes reinterpreted, sometimes applied to problems wholly different from those for which they were first devised--can play a useful role to this end.

Reckase, M. (1996). Test construction in the 1990s: Recent approaches every psychologist should know. *Psychological Assessment, 8*(4), 354-359.

The article summarizes the current state of the art in test construction and contrasts it with previous conceptual models, some of which are wrong or misleading. In particular, new methodologies for item selection and review arc presented as well as current thinking on the specification of technical characteristics of tests.

Test Specifications

Kane, M. (1997). Model-based practice analysis and test specifications. *Applied Measurement in Education, 10*(1), 5-18.

Licensure and certifications decisions are generally based on a chain of inference from the results of a practice analysis to test specifications, to a test, to examinee performance on the test, to a pass-fail decision. This article focuses on the first two steps in this chain of inference: (a) the design of practice analyses and (b) the translation of practice-analysis results into test specifications. The approach taken is to develop a tentative model of practice prior to conducting the practice-analysis and to use this practice model to structure both the data collection and the interpretation of results. The practice model provides a preliminary version of the test specifications that is then empirically tested and redefined in the practice-analysis study. If the practice model is supported by the data, the translation of practice-analysis results into test specifications can be simple and straightforward.

Item Types

Crocker, L. (1997). Assessing content representativeness of performance assessment exercises. *Applied Measurement in Education, 10*(1), 83-95.

Collection of validation evidence for certification presents new challenges for performance assessments when expert judgments of content are used. In particular, the complexity of the exercises, their “new” format, the restricted number of exercises,

maintaining security for memorable exercises, and the need for scoring rubrics create new methodological concerns. The National Broad for Professional Teaching Standards has been addressing these issues in an attempt to establish validity for certification procedures for highly accomplished teachers. This experience provides illustration of how these issues can be addressed.

Stiggins, R. J. (1987). Design and development of performance assessments (Module 1). *Instructional Topics in Educational Measurement* (ITEMS). Madison, WI: NCME. Retrieved March 01, 2009, from http://www.ncme.org/pubs/items/ITEMS_Mod_1_Intro.pdf

Achievement can be, and often is, measured by means of observation and professional judgment. This form of measurement is called performance assessment. Developers of large-scale assessments of communication skills often rely on performance assessments in which carefully devised exercise elicit performance that is observed and judged by trained raters. Teachers also rely heavily on day-to-day observation and judgment. Like other tests, quality performance assessments must be carefully planned and developed to conform to specific rules of test design. This module presents and illustrates those rules in the form of a step-by-step strategy for designing such assessments, through specifications of (a) reasons(s) for assessment, (b) type of performance to be evaluated, (c) exercises that will elicit performance, and (d) systematic rating procedures. General guidelines are presented for maximizing the reliability, validity, and economy of performance assessments.

Test Length

Alsawalmeh, Y. M., & Feldt, L. S. (1999). Testing the equality of independent alpha coefficients adjusted for test length. *Educational and Psychological Measurement, 59*(3), 373-383. Retrieved March 03, 2009, from ERIC database.

In comparing the merits of different item types or test formats, test developers are often called on to make a rigorous comparison of the internal consistency reliabilities of the scores from two competing tests. Such comparisons should be made for tests of equal length in terms of time demands and speededness. If the experimental instruments used in the field trials were not equal in length, estimates of the reliabilities of scores on equal-length tests can be obtained by applying the Spearman-Brown formula. In this article, the authors develop a statistical test for the hypothesis that $\alpha_1' = \alpha_2'$, where α_i' is the extrapolated value of Cronbach's alpha reliability for Test i. The test is shown to exercise tight control of Type I error.

Hambleton, R. K., Mills, C. N., & Simon, R. (1983). Determining the lengths for criterion-referenced tests. *Journal of Educational Measurement, 20*(1), 27-38. Retrieved February 24, 2009, from PsychNET database.

A computer program was developed for determining test length. Item response theory and computer based simulation are used to facilitate consideration and determination made from the process. Five factors used in the computer simulation which affect the reliability and validity are test length, statistical characteristics of the item pool, method of item selection choice of cut-off score, and domain score distribution. Homogenous item pools yielded more consistent results for shortened tests. Decision consistency with heterogeneous item pools was improved with test length, advantages were found for parallel forms. Discriminating items were found to improve decision consistency for all test lengths.

Test Score Use and Interpretations

National Association for College Admission Counseling (NACAC). (2008, September). *Report of the commission on the use of standardized tests in undergraduate admissions.* Arlington, VA: Author. Retrieved March 25, 2009, from http://www.nacacnet.org/PublicationsResources/Marketplace/Pages/TestingCommissionReport.aspx.

To address increasing public concern about standardized admission tests and their greater importance in undergraduate admission over the past decade, NACAC convened a Commission on the Use of Standardized Tests in Undergraduate Admission. After a year-long study, the commission released a final report highlighting its findings.

2. Item Development

Multiple-Choice Items

Albanese, M. (1993). Type K and other complex multiple-choice items: An analysis of research and item properties. *Educational Measurement: Issues and Practice, 12*(1), 28-33.

> Reviews evidence regarding the recommendation to avoid the use of complex multiple choice (CMC) and the similar Type K(TK) test items. It is apparent that TK items have clueing that increases scores, decreases reliability, and may help poorer examinees at the expense of better examinees. TK items are also more frequently deleted on key verification, suggesting that they are more likely to contain identifiable flaws that reflect on the content validity of such scores. With the consideration of added cost associated with higher deletion rates, the use of TK items should be avoided. Evidence for or against the more general CMC format is less clear because there are very few studies of their properties, and the potential number of variations is quite large. The general form of the CMC may hold some promise, but it rapidly becomes quite complex and adds time to the testing situation.

Ascalon, M., Meyers, L., Davis, B., & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education, 20*(2), 153-170.

> This article examined two item-writing guidelines: the format of the item stem and homogeneity of the answer set. Answering the call of Haladyna, Downing, and Rodriguez (2002) for empirical tests of item writing guidelines and extending the work of Smith and Smith (1988) on differential use of item characteristics, a mock multiple-choice driver's license examination was administered to high school students with items having item stems that were either open-ended or in question form and with distractors structured to be either similar or dissimilar to the correct answer. Analyses at the test level indicated that the similarly structured distracters raised the mean difficulty level by .12. No effect was found for item-stem format. Differential item function analyses on each of the test items further supported the effect of distractor similarity on test performance. Implications of this study for item writing and standard setting, as well as implications for future research, are discussed.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education, 10*(2), 133-143.

The purpose of this research was to study the effects of violations of standard multiple-choice item writing principles on test characteristics, student scores, and pass–fail outcomes. Four basic science examinations, administered to year-one and year-two medical students, were randomly selected for study. Test items were classified as either standard or flawed by three independent raters, blinded to all item performance data. Flawed test questions violated one or more standard principles of effective item writing. Thirty-six to sixty-five percent of the items on the four tests were flawed. Flawed items were 0–15 percentage points more difficult than standard items measuring the same construct. Over all four examinations, 646 (53%) students passed the standard items while 575 (47%) passed the flawed items. The median passing rate difference between flawed and standard items was 3.5 percentage points, but ranged from −1 to 35 percentage points. Item flaws had little effect on test score reliability or other psychometric quality indices. Results showed that flawed multiple-choice test items, which violate well established and evidence-based principles of effective item writing, disadvantage some medical students. Item flaws introduce the systematic error of construct-irrelevant variance to assessments, thereby reducing the validity evidence for examinations and penalizing some examinees.

Downing, S. M., Baranowski, R. A., Grosso, L. J., & Norcini, J. J. (1995). Item type and cognitive ability measured: The validity evidence for multiple true-false items in medical specialty certification. *Applied Measurement in Education*, *8*(2), 187-197.

This study compared the criterion-related validity evidence and other psychometric characteristics of multiple-choice (MCQ) and multiple true-false (MTF) items in medical specialty certifying examinations in internal medicine and its subspecialties. Results showed that MTF items were more reliable than MCQs and that the format scores were highly correlated. However, MCQs were more highly correlated with an independent performance measure than were MTF items. MTF items were classified primarily as measuring knowledge rather than synthesis or judgment. These results may have implications for examination construction, especially if criterion-related validity evidence is important.

Frisbie, D. A. (1992). The multiple true-false item format: A status review. *Educational Measurement: Issues and Practice, 11*(4), 21-26.

The literature related to multiple true-false (MTF) items is reviewed to distinguish it from other related formats. A synthesis of past research findings summarizes MTF psychometric characteristics. In general, MTF tests appear efficient and reliable, although examinees perceive MTF items as harder than multiple choice items. A research agenda for the MTF item is proposed.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

A taxonomy of 31 multiple-choice item-writing guidelines was validated through a logical process that included two sources of evidence: the consensus achieved from reviewing what was found in 27 textbooks on educational testing and the results of 27 research studies and reviews published since 1990. This taxonomy is mainly intended for classroom assessment. Because textbooks have potential to educate teachers and future teachers, textbook writers are encouraged to consider these findings in future editions of their textbooks. This taxonomy may also have usefulness for developing test items for large-scale assessments. Finally, research on multiple-choice item writing is discussed both from substantive and methodological viewpoints.

Knowles, S. L., & Welch, C. A. (1992). A meta-analytic review of item discrimination and difficulty in multiple-choice items using "none-of-the-above." *Educational and Psychological Measurement, 52*, 571-578.

The research on using "none-of-the-above" (N-O-T-A) as a test item option continues to find contradictor results on item difficulty and item discrimination. The argument for not using N-O-T-A is that this option is considered more difficult and less discriminating than items with one correct format (regular). The purpose of this study was to conduct a meta-analysis on the difficulty and discrimination of the N-O-T-A test option. A total of twelve articles yielding 20 effect sizes was examined for difficulty, and a total of seven studies yielding eleven effect sized was examined for discrimination. The meta-analysis indicated non-significant effect sized of .01 for discrimination and -.17 for item difficulty. These findings indicate that using "none-of-the-above" as a test item does not results in items of lesser quality than in items not using this option.

Moreno, R., Martínez, R., & Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology, 2*(2), 65-72.

The rigorous construction of items constitutes a field of great current interest for psychometric researchers and practitioners. In previous studies we have reviewed and analyzed the existing guidelines for the construction of multiple-choice items. From this review emerged a new proposal for guidelines that is now, in the present work, subjected to empirical assessment. This assessment was carried out by users of the guidelines and by experts in item construction. The results endorse the proposal for the new guidelines presented, confirming the advantages in relation to their simplicity and efficiency, as well as permitting identification of the difficulties involved in drawing up and organizing

some of the guidelines. Taking into account these results, we propose a new, refined set of guidelines that constitutes a useful, simple, and structured instrument for the construction of multiple-choice items.

Roberts, D. M. (1993). An empirical study on the nature of trick test questions. *Journal of Educational Measurement, 30*(4), 331-344.

This study attempted to better define trick questions and see if students could differentiate between trick and not-trick questions. Phase 1 elicited definitions of trick questions so as to identify essential characteristics. Seven components were found. Phase 2 obtained ratings to see which components of trick questions were considered to be most crucial. The intention of the item constructor and the fact that the questions had multiple correct answers received highest ratings from students. Phase 3 presented a collection of statistics items, some of which were labeled on an a priori basis as being trick or not-trick. The analysis indicated that examinees were able to statistically differentiate between trick and not-trick items, but the difference compared to chance was small. Not-trick items were more successfully sorted than trick items, and trick items that were classified as intentional were sorted about as well as nonintentional items. Evidence seems to suggest that the concept of trickiness is not as clear as some test construction textbook authors suggest.

Constructed-Response Items

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289-303.

Issues pertaining to the quality of performance assessments are discussed. Traditional concepts of reliability and validity are important to performance tasks in that they help to establish the contexts in which such measures can be appropriately used and to create caveats for interpretation of results. Examples, both historical and contemporary, show a remarkable degree of consistency in the characteristics of data from human judgments of performance, data that bear directly on matters of trustworthiness and correctness of inferences from samples of complex performances. In particular, direct assessments of complex performance do not typically generalize from one task to another and thus require careful sampling of tasks to secure an acceptable degree of score reliability and validity for most uses. These observations suggest the pressing need for greater quality control in the design and execution of performance assessments. If such assessments are to have lasting effects on instruction and learning, then their technical properties must be understood and appreciated by the developer and practitioner alike.

Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., Dutka, S., & Katzaroff, M. (2001). The importance of providing background information on the structure and scoring of performance assessments. *Applied Measurement in Education, 13*(1), 1-34.

The purpose of this study was to examine the effects of providing students with background information about the structure and scoring of mathematics performance assessments (PAs). Stratifying by grade, we randomly assigned 16 grade 2 through grade 4 classrooms to 2 conditions. In one condition, 187 students took an initial PA, received a brief orientation on the structure and scoring of PAs, and then took a 2nd, alternate-form PA. In the other condition, 182 students took 2 alternate-form PAs with no intervening orientation. Analyses of variance revealed that students' prior achievement histories mediated the effects of the test-wiseness training on the change between PA trials: Effects were statistically significant and dramatic for above- and at-grade level students but not for below-grade level students. Implications of valid assessment within high-stakes annual testing programs are discussed.

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

We determined the recommendations for preparing and scoring constructed-response (CR) test items in 25 sources (textbooks and chapters) on educational and psychological measurement. The project was similar to Haladyna's (2004) analysis for multiple-choice items. We identified 12 recommendations for preparing CR items given by multiple sources, with 4 of these given by at least half of the sources; and 13 recommendations for scoring CR items given by multiple sources, with 5 given by at least half of the sources. Many recommendations received minority support or were unique to individual sources. Research is needed both on the effect of the recommendations for measurement properties and the extent to which the recommendations are adopted in practice.

Pomplun, M., & Sundbye, N. (1999). Gender differences in constructed response reading items. *Applied Measurement in Education, 12*(1), 95-109.

In this study, we explored gender differences in answers to constructed response reading items from a state assessment program. Construct related and unrelated factors were identified by a content expert after a review of the literature and a pilot study. Four raters were trained to score the identified features on approximately 500 papers evenly divided across 2 grade levels (250 7th and 250 10th graders) and between genders. These features included correctness of answer, unrelated answers, inaccurate answers, number of words written, and a measure of syntactic complexity. The papers rated for these features had

already been assigned holistic scores by local teachers using a state-provided rubric. The relations between these studied features, holistic scores, objective scores, and gender differences was explored through correlations and regression analyses. The results indicate that number of words written and number of unrelated responses showed significant gender differences, were related to holistic scores, and were significant even when the other studied variables were controlled statistically. Further research is needed to investigate how these features could influence raters differentially for male and female students.

Multiple-Choice and Constructed-Response Items

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55-77.

The effects of test consequences, response formats (multiple choice or constructed response), gender, and ethnicity were studied for the math and science sections of a high school diploma endorsement test. There was an interaction between response format and test consequences: Under both response formats, students performed better under high stakes (diploma endorsement) than under low stakes (pilot test), but the difference was larger for the constructed response items. Gender and ethnicity did not interact with test stakes; the means of all groups increased when the test had high stakes. Gender interacted with format; boys scored higher than girls on multiple-choice items, girls scored higher than boys on constructed-response items.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218.

The effects of testing on cognition bear not only on the meaning of what is measured, but also its consequences. Test item formats vary in their typical cognitive demand and in the range of cognitions they sample. Multiple-choice items, in particular, often elicit low-level cognitive processing whereas constructed-response items more often evoke complex thinking. However, the typical cognitions elicited by-differing response formats are less a reflection of the limitations of those formats than they are of typical use. Even when research demonstrates that item response formats vary in cognitive features, psychometric characteristics, and costs of administration and scoring, policy implications remain ambiguous because of tradeoffs along these dimensions. No one format is appropriate for all purposes and on all occasions. This article presents the state of the cumulative research bearing on the differential cognitive demand of item formats, and explores implications of the extant research for testing practice.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, *40*(2), 163-184.

A thorough search of the literature was conducted to locate empirical studies investigating the trait or construct equivalence of multiple-choice (MC) and constructed-response (CR) items. Of the 67 studies identified, 29 studies included 56 correlations between items in both formats. These 56 correlations were corrected for attenuation and synthesized to establish evidence for a common estimate of correlation (true-score correlations). The 56 disattenuated correlations were highly heterogeneous. A search for moderators to explain this variation uncovered the role of the design characteristics of test items used in the studies. When items are constructed in both formats using the same stem (stem equivalent), the mean correlation between the two formats approaches unity and is significantly higher than when using non-stem-equivalent items (particularly when using essay-type items). Construct equivalence, in part, appears to be a function of the item design method or the item writer's intent. (PsycINFO Database Record (c) 2008 APA, all rights reserved)

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment, 9*(1&2), 61-78.

Differential item function (DIF) analyses are a routine part of the development of large-scale assessments. Less common are studies to understand the potential sources of DIF. The goals of this study were (a) to identify gender DIF in a large-scale science assessment and (b) to look for trends in the DIF and non-DIF items due to content, cognitive demands, item type, item text, and visual-spatial or reference factors. To facilitate the analyses, DIF studies were conducted at 3 grade levels and for 2 randomly equivalent forms of the science assessment at each grade level (administered in different years). The DIF procedure itself was a variant of the “Standardization procedure” of Dorans and Kulick (1986) and was applied to very large sets of data (6 sets of data, each involving 60,000 students). It has the advantages of being easy to understand and explain to practitioners. Several findings emerged from the study that would be useful to pass on to test development committees. For example, where there was DIF in science items, MC items tended to favor male examinees and OR items tended to favor female examinees. Compiling DIF information across multiple grades and years increases the likelihood that important trends in the data will be identified and that item writing practices will be informed by more than anecdotal reports about DIF.

3. Test Assembly & Evaluation

Weighting of Various Item Types

Feldt, L. S. (2004). Estimating the reliability of a test battery composite or a test score based on weighted item scoring. *Measurement and Evaluation in Counseling and Development, 37*(3), 184-191.

In some settings, the validity of a battery composite or a test score is enhanced by weighting some parts or items more heavily than others in the total score. This article describes methods of estimating the total score reliability coefficient when differential weights are used with items or parts.

Test Evaluation

van Batenburg, T., & Laros, J. A. (2002). Graphical analysis of test items. *Educational Research and Evaluation, 8*(3), 319-333. Retrieved February 15, 2009, from PsychINFO database.

Graphical Item Analysis (GIA) visually displays the relationship between the total score on a test and the response proportions of the correct and false alternatives of a multiple-choice item. The GIA method provides essential and easily interpretable information about item characteristics (difficulty, discrimination and guessing rate). Low quality items are easily detected with the GIA method because they show response proportions on the correct alternative which decrease with an increase of the total score, or display response proportions of one or more false alternatives which do not decrease with an increase of the total score. The GIA method has two main applications. Firstly, it can be used by researchers in the process of identifying items that need to be excluded from further analysis. Secondly, it can be used by test constructors in the process of improving the quality of the item bank. GIA enables a better understanding of test theory and test construction, especially for those without a background in psychometrics. In this sense, the GIA method might contribute to reducing the gap between the abstract world of psychometrics and the practical world of constructors of achievement tests.

Johnstone, C., Altman, J., & Thurlow, M. (2006). *A State Guide to the Development of Universally Designed Assessments.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved March 6, 2009, from www.cehd.umn.edu/NCEO/onlinepubs/StateGuideUD/default.htm

Universal design for assessments is an approach to educational assessment based on principles of accessibility for a wide variety of end users. Elements of universal design include inclusive test population; precisely defined constructs; accessible, non-biased items; tests that are amenable to accommodations; simple, clear and intuitive instructions and procedures; maximum readability and comprehensibility; and maximum legibility. The purpose of this guide is to provide states with strategies for designing tests from the very beginning, through conceptualization and item construction, field-testing, item reviews, statewide operationalization, and evaluation. The objective is to create tests that present an accurate measure of the knowledge and skills of the diverse population of students enrolled in today's public schools. This guide is accompanied by an online supplement, which can be accessed at www.nceo.info/UDmanual/.

4. Computer related Topics

Comparison of Test Formats

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38*(1), 19-49.

When a computerized adaptive testing (CAT) version of a test co-exists with its paper-and-pencil (P&P) version, it is important for scores from the CAT version to be comparable to scores from its P&P version. The CAT version may require multiple item pools for test security reasons, and CAT scores based on alternate pools also need to be comparable to each other. In this paper, we review research literature on CAT comparability issues and synthesize issues specific to these two settings. A framework of criteria for evaluating comparability was developed that contains the following three categories of criteria: validity criterion, psychometric property/reliability criterion, and statistical assumptions/test administration condition criterion. Methods for evaluating comparability under these criteria as well as various algorithms for improving comparability are described and discussed. Focusing on the psychometric properties/reliability criterion, an example using an item pool of ACT Assessment Mathematics items is provided to demonstrate a process for developing comparable CAT versions and for evaluating comparability. This example illustrates how simulations can be used to improve comparability at the early stages of the development of a CAT. The effects of different specifications of practical constraints, such as content balancing and item exposure rate control, and the effects of using alternate item pools are examined. One interesting finding from this study is that a large part of incomparability may be due to the change from number-correct score-based scoring to IRT ability estimation-based scoring. In addition, changes in components of a CAT such as exposure rate control, content balancing, test length, and item pool size were found to result in different levels of comparability in test scores.

Computer Adaptive Testing

Glas, C. A., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27*(4), 247-261.

To increase the number of items available for adaptive testing and reduce the cost of item writing, the use of techniques of item cloning has been proposed. An important consequence of item cloning is possible variability between the item parameters. To deal with this variability, a multilevel item response (IRT) model is presented which allows

for differences between the distributions of item parameters of families of item clones. A marginal maximum likelihood and a Bayesian procedure for estimating the hyperparameters are presented. In addition, an item-selection procedure for computerized adaptive testing with item cloning is presented which has the following two stages: First, a family of item clones is selected to be optimal at the estimate of the person parameter. Second, an item is randomly selected from the family for administration. Results from simulation studies based on an item pool from the Law School Admission Test (LSAT) illustrate the accuracy of these item pool calibration and adaptive testing procedures.

Veerkamp, W. J., & Glas, C. A. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Education and Behavioral Statistics, 25*(4), 373-389.

Due to previous exposure of items in computer adaptive testing, items may become known to a substantial portion of examinees. A disclosed item is bound to show drift in the item parameter values. In this paper, it is suggested to use a statistical quality control method for the detection of known items. The method is worked out in detail for 2 item response theory models: 1-PL and 3-PL. Adaptive test data are used to re-estimate the item parameters, and these estimates are used in a test of parameter drift. The method is illustrated in a number of simulation studies, including a power study.

5. Scoring

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., Comfort, K., & Othman, A. R. (1998). Analytic versus holistic scoring of science performance. *Applied Measurement in Education, 11*(2), 121-137.

We conducted 2 studies to investigate the interreader consistency, score reliability, and reader time requirements of 3 hands-on science performance tasks. One study involved scoring the responses of students in Grades 5, 8, and 10 on 3 dimensions ("curriculum standards") of performance. The other study computed scores for each of the 3 parts of the grade 5 and 8 tasks. Both studies used analytic and holistic scoring rubrics to grade responses but differed in the characteristics of these rubrics. Analytic scoring tool much longer but led to higher interreader consistency. Nevertheless, when averaged over all the questions in a task, a student's holistic score was just as reliable as the student's analytic score. There was a very high correlation between analytic and holistic scores after they were disattenuated for inconsistencies among readers. Using 2 readers per answer does not appear to be a cost-effective means for increasing the reliability of task scores.

6. General

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.

From APA website: The *Standards* is written for the professional and for the educated layperson and addresses professional and technical issues of test development and use in education, psychology and employment. This book is a vitally important reference for professional test developers, sponsors, publishers, users, policymakers, employers, and students in education and psychology.

Joint Committee on Testing Practices. (2004). *Code of Fair Testing Practices in Education.* Washington, DC: Author.

The Code of Fair Testing Practices in Education (Code) is a guide for professionals in fulfilling their obligation to provide and use tests that are fair to all test takers. Fairness is a primary consideration in all aspects of testing. Fairness implies many things and is not an isolated concept, it must be considered in all aspects of the testing process. The *Code* applies broadly to testing in education regardless of the mode of presentation. The *Code* addresses the roles of test developers and test users separately. The *Code* is directed primarily at professionally developed, although teachers are encouraged to use the guidelines to help improve their testing practices.

NCME Ad Hoc Committee on the Development of a Code of Ethics. (1995). *Code of professional responsibilities in educational measurement.* Washington, DC: NCME.

NCME is providing this Code as a public service for all individuals who are engaged in educational assessment activities in the hope that these activities will be conducted in a professionally responsible manner. This Code applies to any type of assessment that occurs as part of the educational process (i.e. large-scale and small-scale assessment, teacher-conducted assessments, interest and personality measures). The Code presents the professional responsibilities of those who: 1) Develop Assessments, 2) Market and Sell Assessments, 3) Select Assessments, 4) Administer Assessments, 5) Score Assessments, 6) Interpret, Use, and Communicate Assessment Results, 7) Educate About Assessment, and 8) Evaluate Programs and Conduct Research on Assessments.

7. Testing Program Documents

National Center for Educational Statistics. (2008). *National Assessment of Educational Progress*. Retrieved February 14, 2009, from nces.ed.gov/nationsreportcard.

Many consider the NAEP assessment to be the ideal way to develop an assessment. The NAEP Technical Documentation Website illustrates the development process. Through a navigable website, one can find the organizing framework developed by the National Assessment Governing Board that serves as a blueprint for and clearly delineates the item-developmental process, the scoring process, and much more. The website assumes some knowledge of educational measurement and testing. An overview of NAEP is provided for persons with little exposure to the measurement field.

ACT, INC. (2009). *The ACT TEST*. Retrieved February 14, 2009, from www.act.org/aap/index.html.

ACT, the test, is created by ACT Inc., a non-profit organization. A variety of information is provided for researchers. Information Briefs provide information about educational issues and related ACT programs. ACT Research Reports provide documentation about current and past studies. Fairness Reports describe the reviews and analyses the EXPLORE, PLAN, and ACT tests undergo before administration. Technical Manuals document the developmental process of tests at ACT. Information is provided for Policymakers and for Educators.