

4. Scores and Reporting Contexts

Scales for reporting

Beaton, A. E., & Johnson, E. G. (1992). Overview of the scaling methodology used in the national assessment. *Journal of Educational Measurement*, 29(2), 163-175.

The National Assessment of Educational Progress (NAEP) uses item response theory (IRT) based scaling methods to summarize information in complex data sets. The necessity of global scores or more detailed subscores, creation of developmental scales for different ages, and use of scale anchoring for scale interpretation are discussed. [Authors' abstract]

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 355-386). Westport, CT: American Council on Education/Praeger.

The authors provide a small section in their chapter focused on different types of derived scales (e.g., stanines, age-equivalent scores and age-equivalent scores) for score reporting. In addition, they describe different uses of scores and how the uses impact on the types of information that users might value in reports. They make a strong case for more research on score report development, especially experimental work. [Our abstract]

Haertel, E. H. (1991, November). *TRP analyses of issues concerning within-age versus cross-age scales for the National Assessment of Educational Progress*. Report presented to the National Assessment Governing Board, San Diego, CA.

The National Assessment Governing Board of Educational Progress has recently adopted the position that the National Assessment of Educational Progress (NAEP) should employ within-age scaling whenever feasible. The NAEP Technical Review panel (TRP) has studied the issue at some length, and reports on it in this analysis. The first section reviews the evidence concerning the tenability of the psychometric assumptions underlying cross-age (vertical) scaling, and considers whether NAEP trends or comparisons would appear materially different if within-age scaling were applied to existing NAEP data. The second section reviews the possible implications of a shift to within-age scaling for the design of the NAEP objectives frameworks and exercise pools. The third and final section relates cross-age versus within-age scaling to the substantive interpretations and policy implications supported by NAEP data. The panel concludes that in general, if one accepts the premise that cross-age scales are valid and useful, then NAEP cross-age scales are not technically flawed in any obvious ways. However, analyses suggest that cross-age scale comparisons are largely flawed and unhelpful. Overall, the report supports the recent decision of the National Assessment Governing Board to use within-age scales when feasible. [Author's abstract]

Mislevy, R. (2000, February). *Evidentiary relationships among data-gathering methods and reporting scales in surveys of educational achievement*. Paper presented at the workshop of the Committee on NAEP Reporting Practices: Investigating District-Level and Market-Based Reporting, National Research Council, Washington, DC.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, 17(2), 131-154.

Scale-score reporting is a recent innovation in the National Assessment of Educational Progress (NAEP). With scaling methods, the performance of a sample of students in a subject area or subarea can be summarized on a single scale even when different students have been administered different exercises. This article presents an overview of the scaling methodologies employed in the analyses of NAEP surveys beginning with 1984. The first section discusses the perspective on scaling from which the procedures were conceived and applied. The plausible values methodology developed for use in NAEP scale-score analyses is then described, in the contexts of item response theory and average response method scaling. The concluding section lists milestones in the evolution of the plausible values approach in NAEP and directions for further improvement. [Authors' abstract]

Philips, G. W., Mullis, I. V. S., Bourque, M. L., Williams, P. L., Hambleton, R. K., Owen, E. H., & Barton, P. E. (1993). *Interpreting NAEP scales*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Rogers, T., & Nowicki, D. M. (2009, April). *A comparison of four scoring procedures for high-stakes and low-stakes examinations with mixed item formats*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

The interchangeability of scores yielded by three weighting procedures applied to low-stakes achievement tests and to high-stakes examinations containing both selected response (SR) items and constructed response (CR) items in Language Arts and Mathematics was examined. The three scoring procedures included an unweighted procedure in which scores from the set of SR items and the set of CR items/tasks were added; a weighted procedure in which the CR items were weighted so that the CR and SR items contributed equally; and pattern scoring in which each item was individually weighted. While the different weighting procedures yielded similar score distributions for all four tests at the group level, they were sufficiently dissimilar at the student level to warrant using them interchangeably. Pattern scoring provided the smallest standard errors, particularly at the lower end of the ability distribution. Whereas test stakes was not a factor, subject area may be a factor. Further, difference between the three score distributions suggest that care must be taken in choosing one weighting

procedure over the others in a criterion-referenced situation, especially when a cut-score is set in the tail of the score distribution. [Authors' abstract]

Russell, M. (2000). *Summarizing change in test scores: Shortcomings of three common methods*. ERIC Digest.

This Digest introduces the advantages and disadvantages of three commonly used methods of reporting test score changes: (1) change in percentile rank; (2) scale or raw score change; and (3) percent change. The change in percentile rank method focuses on the increase or decrease of the mean percentile ranking for a group of students. This method has two main problems. The first is that calculating the mean percentile rank based on an individual's percentile ranks can provide an inaccurate estimate of a group's mean performance. The second is that, because of unequal intervals separating percentile ranks, changes in percentile ranks represent different amounts of growth at each point on the scale. A second method is scale or raw score change. The main drawback to this method is that when raw scores are used to determine change, it is difficult to compare change across tests with different score ranges. A third approach, that of reporting percent change, causes further distortion. Resulting in a statistic that is difficult to interpret and misleading. All of these methods should be avoided when summarizing change in test scores. A separate Digest suggests better ways to summarize changes. [Author's abstract]

Way, W. D., Forsyth, R. A., & Ansley, T. N. (1989). IRT ability estimates from customized achievement tests without representative content sampling. *Applied Measurement in Education*, 2(1), 15-35.

Examines the effects of using item response theory (IRT) ability estimates based on customized tests that were formed by selecting specific content areas from a nationally standardized achievement test. Tendency of ability estimates and estimated national percentile ranks based on the content-customized tests in school samples to be systematically higher than those based on the full tests. [Author's abstract]

Achievement levels

Crone, C., Zhang, Y., & Kubiak, A. (2006, April). *Cross-validation of proficiency levels for a large scale English language assessment test*. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Hambleton, R. K. (1998). Enhancing the validity of NAEP achievement level score reporting. In M. L. Bourque (Ed.), *Proceedings of the Achievement Levels Workshop* (pp. 77-98). Washington, DC: National Assessment Governing Board.

Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., Nellhaus, J., Reckase, M., Rindone, D., van der Linden, W. J., & Zwick, R.

(2000). A response to “Setting Reasonable and Useful Performance Standards” in the National Academy of Sciences: Grading the nation's report card. *Educational Measurement: Issues and Practice*, 19(2), 5-14.

Responds to a negative evaluation of the National Assessment of Educational Progress (NAEP) by the National Academy of Sciences (NAS) and asserts that a review of the evidence for the NAEP performance standards indicates that there is support for the current approach to NAEP standard setting. Considers the scholarship of the NAS evaluation inadequate. [Authors’ abstract]

Hambleton, R. K., & Slater, S. C. (1995). Using performance standards to report national and state assessment data: Are the reports understandable and how can they be improved? *Proceedings of the Joint Conference on Standard-Setting for Large-Scale Assessments* (pp. 325-343). Washington, DC: NCES.

Considerable evidence suggests that policy-makers, educators, the media, and the public do not understand national and state test results. The problems appear to be two-fold: the scales on which scores are reported seem confusing, and the report forms themselves are often too complex for the intended audiences. This paper addresses two topics. The first is to make test-score reporting scales more meaningful for policymakers, educators, and the media. Of particular importance in work on the National Assessment of Educational Progress (NAEP) was the use of performance standards in score reporting. The second topic is the actual report forms that are used to communicate results. Results from a recent interview study with 60 participants using the Executive Summary of the 1992 NAEP Mathematics Assessment were used to highlight problems in score reporting and to suggest guidelines for improvement. The burden is on the reporting agency to ensure that reporting scales are meaningful and that reported scales are valid for the recommended uses. [Authors’ abstract]

Koretz, D. M., & Deibert, E. (1995/1996). Setting standards and interpreting achievement: A cautionary tale from the National Assessment of Educational Progress. *Educational Assessment*, 3(1), 53-81.

Focuses on the establishment of National Assessment of Educational Progress NAEP on clear performance standards for students in the U.S. Presentation of 1990 NAEP mathematics assessment; Basis of NAEP scale on scoring; Types of characterization of student performance. [Authors’ abstract]

Linn, R. L. (1998). Validating inferences from National Assessment of Educational Progress achievement-level reporting. *Applied Measurement in Education*, 11(1), 23-47.

The validity of interpretations of National Assessment of Educational Progress (NAEP) achievement levels is evaluated by focusing on evidence regarding 3 types of discrepancies: (a) discrepancies between standards implied by judgments

of different types of items (e.g., multiple choice vs. short answer or dichotomously scored vs. extended response tasks scored using multipoint rubrics), (b) discrepancies between descriptions of achievement levels with their associated exemplar items and the location of cut scores on the scale, and (c) discrepancies between the assessments and content standards. Large discrepancies of all 3 types raise serious questions about some of the more expansive inferences that have been made in reporting NAEP results in terms of achievement levels. It is argued that the evidence reviewed provides a strong case for making more modest inferences and interpretations of achievement levels than have frequently been made. [Author's abstract]

National Research Council of the National Academies. (2005). *Measuring literacy: Performance levels for adults*. Washington DC: Author.

Schulz, E. M., Kolen, M. J., & Nicewander, W. A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement, 23*(4), 347-362.

A new procedure for defining achievement levels on continuous scales was developed using aspects of Guttman scaling and item response theory. This procedure assigns examinees to levels of achievement when the levels are represented by separate pools of multiple-choice items. Items were assigned to levels on the basis of their content and hierarchically defined level descriptions. The resulting level response functions were well-spaced and noncrossing. This result allowed well-spaced levels of achievement to be defined by a common percent-correct standard of mastery on the level pools. Guttman patterns of mastery could be inferred from level scores. The new scoring procedure was found to have higher reliability, higher classification consistency, and lower classification error, when compared to two Guttman scoring procedures. [Authors' abstract]

Williams, B., Gawlick, L., & Li, J. (2009, April). *Comparison of indices of classification based on adaptive tests*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Scale anchoring / item mapping

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*(2), 191-204.

The National Assessment of Educational Progress (NAEP) makes possible comparison of groups of students and provides information about what these groups know and can do. The scale anchoring techniques described in this chapter address the latter purpose. The direct method and the smoothing method of scale anchoring are discussed. [Authors' abstract]

Hambleton, R. K., Sireci, S., & Huff, K. (2008). *Development and validation of enhanced SAT score scales using item mapping and performance category descriptions* (Final Report). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics*, 23(1), 35-56.

A procedure is presented for locating on the latent trait scale the scores (or responses) of items that follow the three-parameter logistic (3PL) and mono-tone partial credit (MPC) models. The procedure is based on a Bayesian updating of the item information and is identical to locating the score at the latent trait value that maximizes the Bock score information. Applications are provided in terms of selecting items or score categories for criterion-referenced interpretation and mapping and analyzing score categories. [Author's abstract]

Huynh, H. (2000, April). *On item mappings and statistical rules for selecting binary items for criterion-referenced interpretation and Bookmark standard settings*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Kolstad, A., Cohen, J., Baldi, S., Chan, T., DeFur, E., & Angeles, J. (1998). *The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard?* Washington, DC: American Institutes for Research.

Ryan, J. M. (2003). *An analysis of item mapping and test reporting strategies*. Greensboro, NC: SERVE.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, 20(2), 15-25.

What is item mapping and how does it aid test score interpretation? Which item mapping technique produces the most consistent results and most closely matches expert opinion? [Authors' abstract]

Domain score / subscore reporting

Bock, R. D. (1997). Domain scores: A concept for reporting the National Assessment of Educational Progress results. In R. Glaser, R. Linn, & G. Bohrnstedt (Eds.), *Assessment in transition: Monitoring the Nation's Educational Progress* (pp. 88-102). Stanford, CA: National Academy of Education.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement, 37*(3), 197-211.

Resampling results with real data for 1,000 test responses and 2,902 young adults show that for unidimensional and multidimensional models the item response theory (IRT) estimator is a more accurate predictor of the domain score than is the classical percent-correct score. [Authors' abstract]

de la Torre, J., & Song, H. (2009, April). *A comparison of four methods of IRT subscore*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Lack of sufficient reliability is the primary impediment for generating and reporting subtest scores. Several methods that are currently available improve estimation of subscores by either incorporating the correlation structure among the subtest abilities or utilizing the examinee's performance on the overall test. This paper conducted a systematic comparison among four subscore methods: the multidimensional scoring, the augmented score, the higher-order item response model and the object performance index (OPI) by examining how sample size, test length, number of subtests or domains and their correlations affect the subtest ability estimation. The correlation-based methods provided similar results, and performed best in multiple short subtests measuring highly correlated abilities. The OPI method performed relatively poorer compared to the other methods in all conditions on both ability estimation and proportion correct scores. Real data analysis further underscores the similarities and differences between the four subscore methods. [Authors' abstract]

Edwards, M. C., & Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow? *Journal of Educational and Behavioral Statistics, 31*(3), 241-259.

This article examines a subscore augmentation procedure. The approach uses empirical Bayes adjustments and is intended to improve the overall accuracy of measurement when information is scant. Simulations examined the impact of the method on subscale scores in a variety of realistic conditions. The authors focused on two popular scoring methods: summed scores and item response theory scale scores for summed scores. Simulation conditions included number of subscales, length (hence, reliability) of subscales, and the underlying correlations between scales. To examine the relative performance of the augmented scales, the authors computed root mean square error, reliability, percentage correctly identified as falling within specific proficiency ranges, and the percentage of simulated individuals for whom the augmented score was closer to the true score than was the nonaugmented score. The general findings and limitations of the study are discussed and areas for future research are suggested. [Authors' abstract]

Gessaroli, M. E. (2004, April). *Using hierarchical multidimensional item response theory to estimate augmented subscores*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Haberman, S. J. (2008). *Subscores and validity* (ETS Research Report No. RR-08-64). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33(2), 204–229.

In educational tests, subscores are often generated from a portion of the items in a larger test. Guidelines based on mean squared error are proposed to indicate whether subscores are worth reporting. Alternatives considered are direct reports of subscores, estimates of subscores based on total score, combined estimates based on subscores and total scores, and residual analysis of subscores. Applications are made to data from two testing programs. [Author's abstract]

Haberman, S. J., & Sinharay, S. (2009). *Reporting of subscore using multidimensional item response theory* (ETS Research Report No. RR-09-xx). Princeton, NJ: Educational Testing Service.

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, 62, 79–95.

Recently, there has been an increasing level of interest in reporting subscores for components of larger assessments. This paper examines the issue of reporting subscores at an aggregate level, especially at the level of institutions to which the examinees belong. A new statistical approach based on classical test theory is proposed to assess when subscores at the institutional level have any added value over the total scores. The methods are applied to two operational data sets. For the data under study, the observed results provide little support in favour of reporting subscores for either examinees or institutions. [Authors' abstract]

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation & the Health Professions*, 27(4), 349–368.

Subscores resulting from the administration of high-stakes tests to candidates for credentials in the health professions are desirable for two reasons. First, failing candidates want a profile of performance to plan future remedial studies. Second, training institutions want a profile of performance for their graduates to better evaluate their training. The validity of the interpretation or use of subscores depends on a summative judgment based on a combination of reasoning and empirical analyses, known as validation. We describe this reasoning process and show that with a large credentialing test the validity of any subscore interpretation or use can and should be studied systematically. Validity evidence should be established to support the interpretation and use of subscores that we intend to

report. Some principles arise in this study related to the validity of subscores, and some procedures are proposed to help testing program personnel better validate the use of subscores. [Authors' abstract]

Harris, D. J. (2006, April). *Providing domain scores and national percentile ranks on augmented tests*. Paper presented at the meeting of the National Council of Measurement in Education, San Francisco, CA.

Harris, D. J., & Hanson, B. A. (1991, April). *Methods of examining the usefulness of subscores*. Paper presented at the meeting of the National Council of Measurement in Education, Chicago, IL.

Kahraman, H., & Kamata, A. (2004). Increasing the precision of subscale scores by using out-of-scale information. *Applied Psychological Measurement*, 28(6), 407-426.

In this study, the precision of subscale score estimates was evaluated when out-of-scale information was incorporated. Procedures that incorporated out-of-scale information and only information within a subscale were compared through a series of simulations. It was revealed that more information (i.e., more precision) was always provided for subscale score estimates when out-of-scale information was used. The degree of the information gain depended on the number of out-of-scale items, the magnitude of item discrimination power, and the magnitude of subscale-trait correlation. Also, the accuracy of subscale score estimates was evaluated. Contrary to precision, subscale score estimates were somewhat more biased with out-of-scale information when there were more out-of-scale items and/or when out-of-scale items had high item discrimination power. This tendency was more apparent when the correlation between subscale traits was low. It was concluded that subscale-trait correlation is an important factor to be considered when out-of-scale information is used. [Authors' abstract]

Ling, G. (2009, April). *Report subscores or not? Evaluating subscore reliability and internal test structure*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

The current study evaluated whether to report individual test-takers' subscores of the Major Field Business Test (MFT Business) by analyzing subscores' reliabilities and the internal structure of the test. Reliability analysis found that for each individual student, the observed subscores did not contribute statistically meaningful information beyond the total score of the test. In addition, analysis of internal structure of the MFT Business found a uni-dimensional construct to be present, which also did not support the additional reporting of subscores for each individual student. The relationship between the two analyses was also discussed and an alternate method was recommended for future research. The study concluded that the MFT Business should not report subscores of individual students. [Author's abstract]

Lyrén, P. (2009). Reporting subscores from college admission tests. *Practical Assessment, Research & Evaluation, 14*(4), 3-12. Retrieved April 2, 2009, from <http://pareonline.net/pdf/v14n4.pdf>

The added value of reporting subscores on a college admission test (SweSAT) was examined in this study. Using a CTT-derived objective method for determining the value of reporting subscores, it was concluded that there is added value in reporting section scores (Verbal/Quantitative) as well as subtest scores. These results differ from a study of the SAT I and a study of a basic skills test and thus highlight the need for practitioners and researchers to gather empirical evidence to support the reporting of subscores. The cause of the disparate results seems to be related to differences in the composition of the tests rather than differences in the composition of the examinee groups. [Author's abstract]

McPeck, M., Altman, R., Wallmark, M., & Wingersky, B. C. (1976). *An investigation of the feasibility of obtaining additional subscores on the GRE Advanced Psychology Test* (GRE Board Professional Report No. 74-4P). Princeton, NJ: Educational Testing Service. (ERIC Document No.ED163090)

This study was undertaken to determine whether additional information useful for guidance or placement could be derived from the existing Graduate Record Examinations (GRE) Advanced Psychology Test. The number of subscores currently reported is limited by the high reliability required to make admissions decisions; subscores used only for guidance and placement would not need to meet such a rigorous standard. Subscores based on eight content areas (Personality, Learning, Measurement, Developmental psychology, Social psychology, Physiological and Comparative psychology, Perceptual and Sensory psychology, and Clinical and Abnormal psychology) were identified by the GRE Advanced Psychology Test Committee of Examiners. These experimental subscores, the two currently reported subscores, and the total score were analyzed. Analysis showed that, for most students, additional information about strengths and weaknesses in some of the areas could be obtained. The particular subscores which could provide useful information varied from student to student. This finding was supported by an examination of fifty randomly chosen answer sheets. It was concluded that subscores based on the content areas identified by the Psychology Committee may have potential for providing additional information for purposes of guidance and placement. Subscores based on a factor analysis of the test, however, were judged not to have equivalent potential. [Authors' abstract]

Monaghan, W. (2006). *The facts about subscores* (ETS R&D Connections No. 4). Princeton, NJ: Educational Testing Service. Retrieved January 29, 2009, from http://www.ets.org/Media/Research/pdf/RD_Connections4.pdf

Pei, L. K., Kim, W., & Roussos, L. (2009, April). *Comparison of raw score and diagnostic model-based methods for profile analysis*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

The U.S. government's [No Child Left Behind \(NCLB\) Act of 2001](#) states that all children should be assessed every year to determine whether they are making adequate academic progress, and that students should receive diagnostic reports that allow teachers to address their specific academic needs. Clearly, the quality of test interpretation is crucial to appropriate instructional planning, diagnostic assessment, and educational placement. Profile analysis is one of the most popular test interpretation methods. Profile analysis refers to the determination of cognitive strengths and weaknesses to assist in diagnostic intervention decisions. Pfeiffer, Reddy, Kletzel, Schmelzer, and Boyer (2001) reported that 89% of school psychologists used subtest profile analysis, and 70% of them ranked profile analysis as the most beneficial feature of the Wechsler Intelligence Scale for Children (WISC-III; Wechsler, 1991). The WISC-III manual endorses using profile analysis in classification, stating that "[subtest scatter] variability is frequently considered as diagnostically significant". (p. 177) Due to the popularity of profile analysis in intelligence testing and its importance in educational placement decisions, it is critical to derive profiles in a methodologically rigorous way. Individual student profiles can be defined as an examinee's set of subtest scores on a test battery, such as WISC-III. Other commonly used methods to derive profiles include argument scores (Bock, Thissen & Zimowski, 1997), latent class analysis (Lazarsfeld, 1950) and the fusion model (Roussos, DiBello, Stout, Hartz, Henson, & Templin, 2007). Among these methods, the fusion model not only links students' test score to a statistical model but also links test score to cognitive theory. This paper describes an empirical study comparing profiles based on raw subscores to those based on mastery probability from the fusion model. [Authors' introduction]

Pommerich, M., Nicewander, W. A., & Hanson, B. (1999). Estimating average domain scores. *Journal of Educational Measurement*, 36(3), 199-216.

A simulation study was performed to determine whether a group's average percent correct in a content domain could be accurately estimated for groups taking a single test form and not the entire domain of items. Six Item Response Theory (IRT)-based domain score estimation methods were evaluated, under conditions of few items per content area per form taken, small domains, and small group sizes. The methods used item responses to a single form taken to estimate examinee or group ability; domain scores were then computed using the ability estimates and domain item characteristics. The IRT-based domain score estimates typically showed greater accuracy and greater consistency across forms taken than observed performance on the form taken. For the smallest group size and least number of items taken, the accuracy of most IRT-based estimates was questionable; however, a procedure that operates on an estimated distribution of group ability showed promise under most conditions. An appendix discusses

estimating mean group ability using a latent-variable regression model. [Authors' abstract]

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (in press). Comparison of subscores based on classical test theory. *Applied Psychological Measurement*.

Sinharay, S. (2009). *When can subscores be expected to have added value? Results from operational and simulated data* (ETS Research Memorandum). Princeton, NJ: ETS.

Sinharay, S., & Haberman, S. (2008). *Reporting subscores: A survey* (Research Report RM-08-18). Princeton, NJ: Educational Testing Service.

Recently, there has been an increasing level of interest in subscores for their potential diagnostic value. As a result, there is a constant demand from test users for subscores. Haberman (2005) and Haberman, Sinharay, and Puhan (2006) suggested methods based on classical test theory to examine whether subscores provide any added value over total scores. This paper applied the above mentioned methods to recent data sets from a variety of operational tests. The results indicate that subscores provide added value for only a handful of tests. [Authors' abstract]

Sinharay, S., & Haberman, S. J. (2009). How much can we reliability know about what students know? *Measurement: Interdisciplinary Research and Perspectives*, 7(1), 46-49.

The authors reflect on the issues regarding practitioners' use of diagnostic classification models (DCMs). They cite several issues including the lack of studies that demonstrate the validity of the results and information provided by DCMs, and the unreported classification reliability obtained by DCMs. They also provide recommendations on diagnostic scoring for potential DCM users including the sufficiency of reported diagnostic information. [Authors' abstract]

Sinharay, S., Haberman, S., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, 26(4), 21-28.

There is an increasing interest in reporting subscores, both at examinee level and at aggregate levels. However, it is important to ensure reasonable subscore performance in terms of high reliability and validity to minimize incorrect instructional and remediation decisions. This article employs a statistical measure based on classical test theory that is conceptually similar to the test reliability measure and can be used to determine when subscores have any added value over total scores. The usefulness of subscores is examined both at the level of the examinees and at the level of the institutions that the examinees belong to. The suggested approach is applied to two data sets from a basic skills test. The results

provide little support in favor of reporting subscores for either examinees or institutions for the tests studied here. [Authors' abstract]

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education, 17*(2), 89-112.

The valid provision of subscores from an item response theory-based test implies a multidimensional test structure. Assuming, in the construction of a new test, that the test features required for a valid and reliable total test score have been specified already, this article describes the resulting subscore performance and the resulting degradation of the total score performance caused by multidimensionality. Subscore and total score error variances for both maximum likelihood and expected a posteriori estimators were determined for a typical test as a function of the test dimensionality (i.e., the number of subscores) and the level of correlation among the subscore abilities. The hit rates for detecting true differences among subscore abilities of practical importance are presented. [Author's abstract]

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287-307.

Probabilistic models with one or more latent variables are designed to report on a corresponding number of skills or cognitive attributes. Multidimensional skill profiles offer additional information beyond what a single test score can provide, if the reported skills can be identified and distinguished reliably. Many recent approaches to skill profile models are limited to dichotomous data and have made use of computationally intensive estimation methods such as Markov chain Monte Carlo, since standard maximum likelihood (ML) estimation techniques were deemed infeasible. This paper presents a general diagnostic model (GDM) that can be estimated with standard ML techniques and applies to polytomous response variables as well as to skills with two or more proficiency levels. The paper uses one member of a larger class of diagnostic models, a compensatory diagnostic model for dichotomous and partial credit data. Many well-known models, such as univariate and multivariate versions of the Rasch model and the two-parameter logistic item response theory model, the generalized partial credit model, as well as a variety of skill profile models, are special cases of this GDM. In addition to an introduction to this model, the paper presents a parameter recovery study using simulated data and an application to real data from the field test for TOEFL® Internet-based testing. [Author's abstract]

Wainer, H., Vevea, J. L., Camacho, F., Reeve III, B. B., Rosa, K., Nelson, L., Swygert, K. A., & Thissen, D. (2000). Augmented scores—"borrowing strengths" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Ed.), *Test scoring* (pp. 343-387). Mahwah, NJ: Lawrence Erlbaum Associates.

The authors introduce the general principles of empirical Bayes estimation, and then use those principles to develop multivariate generalization of T. L. Kelley's

(1927) regressed estimates of true scores. The goal of this development is the computation of reliable estimates of subscores. Topics discussed include: regressed estimates: statistical augmentation of meager information; an observed score approach to augmented scores; and an approach to augmented scores that uses linear combinations of item response theory scale scores. [Authors' abstract]

Yao, L. (2009, April). *Reporting valid and reliable overall score and domain score*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

No Child Left Behind (NCLB, 2002) requires state assessment in both report overall (or composite) score and report domain (or objective) scores. Solutions that not only estimate students' accountability levels, but also provide students and their teachers with useful diagnostic information-in addition to the single "overall" score-are desirable. In practice, overall scores were obtained by simply averaging the domain scores. However, simply averaging the domain scores ignores the fact that different domains have different score points, that scores from those domains are related, and that at different score points, the relationship between overall score and domain score may be different. In order to report reliable and valid overall scores and domain scores, we investigated the performance of three procedures through both real data and simulation data, which are the following: 1) Unidimensional IRT model; 2) Higher Order IRT (HO-IRT) model, simultaneous estimate the overall ability and domain abilities; 3) Multidimensional IRT (MIRT) model to estimate domain abilities, with the maximum information method to obtain the overall ability. Our findings suggest that the MIRT model not only provides reliable domain scores, but also produces a reliable overall score that has the smallest standard error of measurement through use of the maximum information method, without assuming any linear relationship between overall score and domain scores, as the other models do. Suggestions for the conditions, such as the correlation between domains and the number of items needed, were recommended for such reporting purposes. [Author's abstract]

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*(2), 83–105.

Several approaches to reporting subscale scores can be found in the literature. This research explores a multidimensional compensatory dichotomous and polytomous item response theory modeling approach for subscale score proficiency estimation, leading toward a more diagnostic solution. It also develops and explores the recovery of a Markov chain Monte Carlo (MCMC) estimation approach to multidimensional item and ability parameter estimation, as well as subscale proficiency and classification rates. The simulation study presented here used real data-derived parameters from a large-scale statewide

assessment with subscale score information under varying conditions of sample size and correlations between subscales (.0, .1, .3, .5, .7, .9). It was found that to report accurate diagnostic information at the subscale level, the subscales need to be highly correlated, or a multidimensional approach should be implemented. MCMC methodology is still a nascent methodology in psychometrics; however, with the growing body of research, its future looks promising. [Authors' abstract]

Diagnostic score reporting

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, 7(4), 255–278.

Item response theory (IRT) describes the interaction between examinees and items using probabilistic models. One of the underlying assumptions of IRT is that examinees are all using the same skill or same composite of multiple skills to respond to each of the test items. When item response data do not satisfy the unidimensionality assumption, multidimensional item response theory (MIRT) should be used to model the item-examinee interaction. MIRT enables one to model the interaction of items that are capable of discriminating between levels of several different abilities and examinees that vary in their proficiencies on these abilities. In this article graphical MIRT analyses designed to provide better insight into what individual items are measuring as well as what the test as a whole is assessing are presented and discussed. The goal of the article is to encourage testing practitioners to use MIRT as a means to statistically validate the test specifications. [Author's abstract]

Ackerman, T., & Shu, Z. (2009, April). *Using confirmatory MIRT modeling to provide diagnostic information in large scale assessment*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

This paper examines different approaches of using multidimensional item response compensatory models to obtain diagnostic information. In this research a large scale assessment of a mid-western state was used. Specifically, the data that were calibrated in this study came from a fifth grade End-of-Grade (EOG) assessment of reading ability. It contained a total of 73 multiple choice items. According to the test specification manual 55 items were intended to measure reading ability (i.e., the understanding and meaning of words and phrases) and the remaining 18 items were intended to measure comprehension (i.e., understanding the characters and purpose of a passage). In all four different item response theory models ranging from a two-parameter unidimensional model to a three-dimensional bifactor model were fit to the data. Results were analyzed and corresponding mastery vs. non-mastery decisions were made based upon the calibrated results. [Authors' abstract]

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J. (2007). Modeling diagnostic assessment with Bayesian networks. *Journal of Educational Measurement*, 44(4), 341–359.

This paper defines Bayesian network models and examines their applications to IRT-based cognitive diagnostic modeling. These models are especially suited to building inference engines designed to be synchronous with the finer grained student models that arise in skills diagnostic assessment. Aspects of the theory and use of Bayesian network models are reviewed, as they affect applications to diagnostic assessment. The paper discusses how Bayesian network models are set up with expert information, improved and calibrated from data, and deployed as evidence-based inference engines. Aimed at a general educational measurement audience, the paper illustrates the flexibility and capabilities of Bayesian networks through a series of concrete examples, and without extensive technical detail. Examples are provided of proficiency spaces with direct dependencies among proficiency nodes, and of customized evidence models for complex tasks. This paper is intended to motivate educational measurement practitioners to learn more about Bayesian networks from the research literature, to acquire readily available Bayesian network software, to perform studies with real and simulated data sets, and to look for opportunities in educational settings that may benefit from diagnostic assessment fueled by Bayesian network modeling. [Authors' abstract]

Bolt, D. (2007). The present and future of IRT-based cognitive diagnostic models (ICDMs) and related methods. *Journal of Educational Measurement*, 44(4), 377-383.

As the goals of educational assessment evolve from the strictly evaluative to the diagnostically useful, so also evolve the statistical methods used to build, validate, and interpret educational tests. The methods discussed in this special issue all approach diagnosis in an item response theory (IRT) related way, with models that are parameterized at the item level and that extract information from individual item responses. Clearly, their most distinguishing feature is their more complex, multidimensional representation of examinee proficiency. This representation can be built directly into an item response model (as seen in most clearly in Almond, DiBello, Moulder, & Zapata-Rivera, 2007; Henson, Templin, & Douglas, 2007; Roussos, Templin, & Henson, 2007; Stout, 2007) or else it can provide a framework for interpreting (residual) patterns in item responses (as is seen in Gierl, 2007).

The complexity of the proficiency space introduces corresponding complexities into the statistical modeling and score reporting aspects of diagnosis. A high level of expert judgment is needed in formulating appropriate models. One of the primary challenges in implementing IRT-based cognitively diagnostic model (ICDMs) requires determining which aspects of the modeling process should be constrained through expert judgment and which can and should be informed by observed item response data. The vast array of psychometric models now

available for diagnosis and the different ways they handle these complexities (e.g., how many levels for each skill, how do skills interact, how does skill mastery translate to item performance, etc.) make model selection a central issue. At the same time, it can be challenging to compare models according to goodness of fit due to the many other aspects within each model that must be informed by experts (e.g., entries of the item-by-skill Q matrix, structure of the proficiency space, etc). Data-driven model re-specification is often messy.

Collectively, the papers presented in this Special Issue provide a comprehensive overview of the state of the art in IRT-based diagnosis. While all emphasize a common end-goal of examinee diagnosis, the process by which this is achieved and the balance of data-driven and expert-driven decision making used along the way also introduce important differences. [Author's abstract]

Clauser, B. E., Subhiyah, R., Nungester, R. J., Ripkey, D., Clyman, S. G., & McKinley, D. (1995). Scoring a performance-based assessment by modeling the judgments of experts. *Journal of Educational Measurement*, 32(4), 397-415.

Performance assessments typically require expert judges to individually rate each performance. These results in a limitation in the use of such assessments because the rating process may be extremely time consuming. This article describes a scoring algorithm that is based on expert judgments but requires the rating of only a sample of performances. A regression-based policy capturing procedure was implemented to model the judgment policies of experts. The data set was a seven-case performance assessment of physician patient management skills. The assessment used a computer-based simulation of the patient care environment. The results showed a substantial improvement in correspondence between scores produced using the algorithm and actual ratings, when compared to raw scores. Scores based on the algorithm were also shown to be superior to raw scores and equal to expert ratings for making pass/fail decisions which agreed with those made by an independent committee of experts. [Authors' abstract]

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3), 333-353.

Higher-order latent traits are proposed for specifying the joint distribution of binary attributes in models for cognitive diagnosis. This approach results in a parsimonious model for the joint distribution of a high-dimensional attribute vector that is natural in many situations when specific cognitive information is sought but a less informative item response model would be a reasonable alternative. This approach stems from viewing the attributes as the specific knowledge required for examination performance, and modeling these attributes as arising from a broadly-defined latent trait resembling the θ of item response models. In this way a relatively simple model for the joint distribution of the attributes results, which is based on a plausible model for the relationship between general aptitude and specific knowledge. Markov chain Monte Carlo algorithms

for parameter estimation are given for selected response distributions, and simulation results are presented to examine the performance of the algorithm as well as the sensitivity of classification to model misspecification. An analysis of fraction subtraction data is provided as an example. [Authors' abstract]

de la Torre, J., & Karelitz, T. M. (2008, March). *When do measurement models produce diagnostic information? An investigation of the assumptions of cognitive diagnostic modeling*. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.

DiBello, L.V. (2002, April). *Skills-based scoring models for the PSAT/NMSQT™*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans.

DiBello, L. V., & Crone, C. (2001, April). *Technical methods underlying the PSAT/NMSQT™ enhanced score report*. Paper presented at the meeting of the National Council on Measurement in Education, Seattle.

DiBello, L.V., & Crone, C. (2001, July). *Enhanced score reporting on a national standardized test*. Paper presented at the International meeting of the Psychometric Society, Osaka, Japan.

DiBello, L. V., Crone, C., Monfils, L., Narcowich, M., & Roussos, L. (2002, April). *Student Profile Scoring*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans.

DiBello, L. V., Stout, W., & Roussos, L. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361-389). Hillsdale, NJ: Erlbaum.

DiBello, L. V., Templin, J., & Henson, R. (2004, June). *Large-scale student profile scoring: Applications to operational tests-next generation TOEFL*. Paper presented at the meeting of the Psychometric Society in Pacific Grove, CA.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56(3), 495-515.

A latent trait model is presented for the repeated measurement of ability based on a multidimensional conceptualization of the change process. A simplex structure is postulated to link item performance under a given measurement condition or occasion to initial ability and to one or more modifiabilities that represent individual differences in change. Since item discriminations are constrained to be equal within a measurement condition, the model belongs to the family of multidimensional Rasch models. Maximum likelihood estimators of the item parameters and abilities are derived, and an example provided that shows good

- recovery of both item and ability parameters. Properties of the model are explored, particularly for several classical issues in measuring change. [Author's abstract]
- Embretson, S. E. (1997). Multicomponent latent trait models. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305-322). New York, NY: Springer-Verlag.
- Gierl, M., Alves, C., Gotzmann, A., Roberts, M. (2009, April). *Using judgments from content specialists to develop cognitive models for diagnostic assessments*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.
- Henson, R., & Douglas, J. (2003). *Using cognitive diagnostic models for development of efficient sumscores*. Princeton, NJ: Educational Testing Service External Research Group Technical Report.
- Henson, R., & Templin, J. (2004). *Creating a proficiency scale with models for cognitive diagnosis*. Princeton, NJ: Educational Testing Service External Research Group Technical Report.
- Henson, R., Templin, J., & Douglas, J. (2007). Using efficient model based sum-scores for conducting skills diagnoses. *Journal of Educational Measurement*, 44(4), 361–376.

Consider test data, a specified set of dichotomous skills measured by the test, and an IRT cognitive diagnosis model (ICDM). Statistical estimation of the data set using the ICDM can provide examinee estimates of mastery for these skills, referred to generally as attributes. With such detailed information about each examinee, future instruction can be tailored specifically for each student, often referred to as formative assessment. However, use of such cognitive diagnosis models to estimate skills in classrooms can require computationally intensive and complicated statistical estimation algorithms, which can diminish the breadth of applications of attribute level diagnosis. We explore the use of sum-scores (each attribute measured by a sum-score) combined with estimated model-based sum-score mastery/nonmastery cutoffs as an easy-to-use and intuitive method to estimate attribute mastery in classrooms and other settings where simple skills diagnostic approaches are desirable. Using a simulation study of skills diagnosis test settings and assuming a test consisting of a model-based calibrated set of items, correct classification rates (CCRs) are compared among four model-based approaches for estimating attribute mastery, namely using full model-based estimation and three different methods of computing sum-scores (simple sum-scores, complex sum-scores, and weighted complex sum-scores) combined with model-based mastery sum-score cutoffs. In summary, the results suggest that model-based sum-scores and mastery cutoffs can be used to estimate examinee attribute mastery with only moderate reductions in CCRs in comparison with the full model-based estimation approach. Certain topics are mentioned that are

currently being investigated, especially applications in classroom and textbook settings. [Authors' abstract]

Henson, R., Templin, J., & Irwin, P. (2009, April). *Ancillary random effects: A way to obtain diagnostic information from existing large scale tests*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

The purpose of this paper is to expand the Log-Linear Cognitive Diagnosis Model (LCDM) (Henson, Templin, and Willse, 2008) to also include and estimate continuous ability measures. These continuous abilities can be defined as effects that are related to an examinee's response for particular items (or all items depending on the test). In many ways, the continuous abilities will function in a similar way as random effects in a mixed model. Thus, the ancillary dimensions will account for dependencies or nuisance dimensions in the data, which allow a more direct assessment of the attributes of interest. After defining this model an illustrative example will be presented using a large scale state assessment where, first, the initial challenges of fitting the LCDM will be discussed and then compared to the LCDM with a single ancillary dimension. By fitting the LCDM with a single continuous dimension one application of the extended new model will be presented using a categorical bi-factor model where the ancillary dimension represents the general factor and the attributes represent the specific factors. [Authors' introduction]

Ho, A., Zapata, D., & Templin, J. (2004, June). *Large-scale student profile scoring: Fast classification and other operational issues for large scale testing*. Paper presented at the meeting of the Psychometric Society in Pacific Grove, CA.

Huff, K. L. (2003). *An item modeling approach to descriptive score reports*. Unpublished doctoral dissertation, University of Massachusetts Amherst, School of Education.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60). New York, NY: Cambridge University Press.

In this chapter, we explore the nature of the demand for cognitive diagnostic assessment (CDA) in K-12 education and suggest that the demand originates from two sources: assessment developers who are arguing for radical shifts in the way assessments are designed, and the intended users of large-scale assessments who want more instructionally relevant results from these assessments. We first highlight various themes from the literature on CDA that illustrate the demand for CDA among assessment developers. We then outline current demands for diagnostic information from educators in the United States by reviewing results from a recent national survey we conducted on this topic. Finally, we discuss

some ways that assessment developers have responded to these demands and outline some issues that, based on the demands discussed here, warrant further attention. [Authors' abstract]

Ketterlin-Geller, L., & Yovanoff, P. (2009, April). *Model comparisons: Fitting cognitive diagnostic models to data*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Lu, Y., & Smith, R. (2009, April). *An alternative method to estimate cluster performance of proficient students on a large scale state assessment*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Almost every state assessment reports cluster scores that reflect performance on different content standards that the test is designed to measure. Although the test blueprints usually specify distributions of items at the individual standard level, for reporting purposes, the content for each test is aggregated across standards into subcontent areas, referred to as "reporting clusters." A student's cluster score is commonly reported as the percentage of items answered correctly out of all items in the cluster. Unlike the total test scores, cluster scores are not equated. Therefore, in order to provide students, parents and educators with more useful information, the cluster scores at the individual or group level need to be compared to some kind of criterion measure or population performance. This paper investigates how this criterion measure is provided on one state assessment and suggests an alternative method to obtain the estimate of the measure. [Authors' introduction]

Luecht, R. M. (2003, April). *Applications of multidimensional diagnostic scoring for certification and licensure tests*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

This paper discusses two topics related to *diagnostic score reporting* for credentialing examinations. The first deals with various ways to compute subscores for credentialing examinations. The second addresses some pertinent factors to consider when presenting *diagnostic* results. To illustrate these issues, a sample set of subscores is used. This set was derived from a certification test that provides pass/fail decisions on multiple sections. There are a number of ways to compute *diagnostic* subscores for competency areas; the paper discusses four approaches. A simulation study using these approaches shows the complexity of choosing a scoring model for multidimensional subscore *reporting*. The decision to use a given method to compute *diagnostic scores* should blend technical sophistication with operational needs. There is very little research literature on presenting *scores*, but there are a number of techniques from which to choose, including *score* tables, profile plots, and narrative text. Producing high quality *score* reports is feasible even for relatively small testing programs. [Author's abstract]

Luecht, R. M. (2007). Using information from multiple-choice distractors to enhance cognitive-diagnostic score reporting. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 319-340). New York, NY: Cambridge University Press.

This chapter focuses on data augmentation mechanisms that make use of any measurement information hidden in meaningful distractor patterns for multiple-choice questions (MCQs). Results are presented from an empirical study that demonstrates that there are reasonable consistencies in MCQ distractor response patterns that might be detected and possibly exploited for *diagnostic* scoring purposes. [Author's abstract]

Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (2006, April). *Scalability and the development of useful diagnostic scales*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

McGlohen, M. K. (2004). *The application of cognitive diagnosis and computerized adaptive testing to a large-scale assessment*. Unpublished doctoral dissertation, University of Texas at Austin.

Michel, R. S. (2007). *The development of a cognitive model to provide psychometrically sound and useful diagnostic information for a quantitative measure*. Unpublished doctoral dissertation, Fordham University, NY.

Milewski, G. B., Baron, P. A. (2002). *Extending DIF methods to inform aggregate reports on cognitive skills*. Paper presented at the meeting of the National Council of Measurement in Education, New Orleans.

Nichols, P. D. (1994). A framework of developing cognitively diagnostic assessments. *Review of Educational Research*, 64(4), 575-603.

The loosely connected efforts to develop cognitively diagnostic assessments are organized. Assessments have been developed to guide specific instructional decisions. [Author's Abstract]

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.

Norris, S. P., Macnab, J. S., & Phillips, L. M. (2007). Cognitive modeling of performance on diagnostic achievement tests: A philosophical analysis and justification. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 61-84). New York, NY: Cambridge University Press.

To interpret and use achievement test scores for cognitive diagnostic assessment, an explanation of student performance is required. If performance is to be explained, then reference must be made to its causes in terms of students'

understanding. Cognitive models are suited, at least in part, to providing such explanations. In the broadest sense, cognitive models should explain achievement test performance by providing insight into whether it is students' understanding (or lack of it) or something else that is the primary cause of their performance. Nevertheless, cognitive models are, in principle, incomplete explanations of achievement test performance. In addition to cognitive models, normative models are required to distinguish achievement from lack of it. The foregoing paragraph sets the stage for this chapter by making a series of claims for which we provide philosophical analysis and justification. First, we describe the philosophical standpoint from which the desire arises for explanations of student test performance in terms of causes. In doing this, we trace the long-held stance within the testing movement that is contrary to this desire and argue that it has serious weaknesses. Second, we address the difficult connection between understanding and causation. Understanding as a causal factor in human behavior presents a metaphysical puzzle: How is it possible for understanding to cause something else to occur? It is also a puzzle how understanding can be caused. We argue that understanding, indeed, can cause and be caused, although our analysis and argument are seriously compressed for this chapter. Also, in the second section, we show why understanding must be taken as the causal underpinning of achievement tests. Third, we examine how cognitive models of achievement might provide insight into students' understanding. This section focuses on what cognitive models can model. Fourth, we discuss what cognitive models cannot model, namely, the normative foundations of achievement, and refer to the sort of normative models that are needed in addition. Finally, we provide an overall assessment of the role and importance of cognitive models in explaining achievement test performance and supporting diagnostic interpretations. [Authors' abstract]

Park, C., & Bolt, D. (2007). *Application of multilevel IRT to investigate cross-national skill profiles on the TIMSS assessment*. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Roussos, L. (1994). *Summary and review of cognitive diagnosis models*. Unpublished manuscript, University of Illinois, Urbana-Champaign, The Statistical Laboratory for Educational and Psychological Measurement.

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement*, 44(4), 293–311.

This article describes a latent trait approach to skills diagnosis based on a particular variety of latent class models that employ item response functions (IRFs) as in typical item response theory (IRT) models. To enable and encourage comparisons with other approaches, this description is provided in terms of the main components of any psychometric approach: the ability model and the IRF structure; review of research on estimation, model checking, reliability, validity, equating, and scoring; and a brief review of real data applications. In this manner

the article demonstrates that this approach to skills diagnosis has built a strong initial foundation of research and resources available to potential users. The outlook for future research and applications is discussed with special emphasis on a call for pilot studies and concomitant increased validity research. [Authors' abstract]

Rudner, L. M., & Talento-Miller, E. (2007, April). *Diagnostic testing using decision theory*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Ruiz-Primo, M., Shavelson, R. J., Li, M., & Schultz, S. E. (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment, 7*(2), 99-141.

The emergence of alternative forms of achievement assessment and the corresponding claims that they measure "higher order thinking" rouse the need to examine their cognitive validity. In this article, we provide a framework for examining cognitive validity claims that includes conceptual and empirical analyses and use it to evaluate the validity of a "connected understanding" *interpretation* of 3 concept-mapping techniques: (a) construct-a-map from scratch, in which students constructed a map using concepts provided; (b) fill-in-the-nodes, in which students filled in a 12-blank-node skeleton map with concepts provided; and (c) fill-in-the-lines, in which students filled in a 12-blank-line skeleton map with a description of the relation provided for each pair of connected concepts. The first technique imposes little structure on the students (low-directedness), whereas the other 2 techniques are much more structured (high-directedness). The framework focuses on the analysis of the mapping tasks' intended demands (conceptual analysis), and the tasks' correspondence with inferred cognitive activities and performance *scores* (empirical analyses). To infer cognitive activities, we examined respondents' (teachers, expert students, and novice students) concurrent and retrospective verbalizations in performing the mapping tasks and compared the directedness of the mapping tasks, the characteristics of verbalization, and the *scores* obtained across techniques. We concluded that the framework allowed us to determine that (a) the 3 mapping techniques provided different pictures of students' knowledge, and (b) inferred cognitive activities across mapping techniques differed in relation to the directedness of the task. The low-directed technique provided students with more opportunities to reveal their conceptual understanding (explanations and errors) than did the high-directed techniques. [Authors' abstract]

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*(4), 333-352.

Discusses the tree-based approach (TBA) which is used for diagnostic feedback for the SAT I Verbal reasoning test, for proficiency scaling and diagnostic assessment. In depth look at the tree-based theory; Use of tree-based techniques to

determine strategic combinations of skills; Generation of group-level proficiency profiles. [Author's abstract]

Sheehan, K. M., Tatsuoka, K. K., & Lewis, C. (1993). *A diagnostic classification model for document processing skills* (Research Report No. RR-93-39). Princeton, NJ: Educational Testing Service.

This paper introduces a modification to the Rule Space diagnostic classification procedure which allows for processing of response vectors containing missing data. Rule Space is an approach to diagnostic classification which involves characterizing examinees' performances in terms of an underlying cognitive model of generalized problem-solving skills. It has two components: (1) a procedure for determining a comprehensive set of knowledge states, where each state is characterized in terms of a unique subset of mastered skills; and (2) a procedure for classifying examinees into one or another of the specified states. The procedure for determining a comprehensive set of knowledge states is based on the Boolean descriptive function given in Tatsuoka (1991). The procedure for classifying examinees involves comparing examinees' scored response vectors to the patterns expected within each of the specified knowledge states (Tatsuoka, 1983, 1985, and 1987). Missing data is expected to be a common problem for this approach because, although the procedure for determining the comprehensive set of knowledge states requires a large pool of items, the procedure for examinee classification can be performed with smaller (less expensive) item subsets. This approach to diagnostic classification is illustrated with data collected in the Survey of Young Adult Literacy, a nationwide survey of literacy skills conducted by the National Assessment of Educational Progress (NAEP) in 1985. [Authors' abstract]

Sinharay, S., Puhan, G, & Haberman, S. J. (2009, April). *Reporting diagnostic scores: Temptations, pitfalls, and some solutions*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Diagnostic scores are of increasing interest due to their potential remedial and instructional benefit. Naturally, the number of testing programs that report diagnostic scores is on the rise, as are the number of research works on such scores. This paper starts by showing examples of diagnostic subscores reported by operational testing programs. Then this paper provides a discussion of existing psychometric methods for reporting diagnostic scores, followed by a brief review of a method proposed by Haberman (2008) that examines if subscores (that are the simplest form of diagnostic scores and are reported by several testing programs) have added value over the total score. Using results from several operational and simulated data sets, it is demonstrated that it is not straightforward to have diagnostic scores with added value. Some recommendations are made for those interested to report diagnostic scores. [Authors' abstract]

- Stone, C. A., & Lane, S. (2008). *Issues in providing subscale scores for diagnostic information*. Retrieved March 28, 2009, from http://www.ccsso.org/content/PDFs/41_Stone_Lane.pdf
- Stone, C. A., Ye, F., Zhu, X., & Lane, S. (in press). *Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional*. *Applied Measurement in Education*.
- Stout, W. (2007). Skills diagnosis using IRT-based continuous latent trait models. *Journal of Educational Measurement, 44*(4), 313–324.

This article summarizes the continuous latent trait IRT approach to skills diagnosis as particularized by a representative variety of continuous latent trait models using item response functions (IRFs). First, several basic IRT-based continuous latent trait approaches are presented in some detail. Then a brief summary of estimation, model checking, and assessment scoring aspects are discussed. Finally, the University of California at Berkeley multidimensional Rasch-model-grounded SEPUP middle school science-focused embedded assessment project is briefly described as one significant illustrative application. [Author's abstract]

- Tatsuoka, K. K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Mahwah, NJ: Lawrence Erlbaum Associates.
- Tatsuoka, K. K., Birenbaum, M., Lewis, C., & Sheehan, K. (1992). *Proficiency scaling based on attribute characteristic curves* (Technical Report No. RR-92-14-ONR). Princeton, NJ: Educational Testing Service.
- Tatsuoka, K. K., & Hayashi, A. (2001). Statistical method for individual cognitive diagnosis based on latent knowledge state. *Journal of The Society of Instrument and Control Engineers, 40*(8), 561-567 (in Japanese).
- Templin, J., He, X., Roussos, L., & Bolt, D. (2004, April). *Polytomous (graded response) item and polytomous (graded) attribute scoring*. Paper presented at the meeting of the National Council on Measurement in Education in San Diego.
- Templin, J., & Henson, R. (2009, April). *Practical issues in using diagnostic estimates: Measuring the reliability and validity of diagnostic estimates*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Over the past decade, diagnostic classification models (DCMs) have become an active area of psychometric research. Despite their use, however, the reliability of examinee estimates in DCM applications has seldom been reported (Sinaharay &

Haberman, in press). In this paper, a reliability measure for the latent variables of DCMs is defined, emanating from a similar measure from more common psychometric models (e.g., item response models). Using theoretical and simulation based results, we show how DCMs uniformly provide greater reliability than IRT models for tests of the same length, a result that is a consequence of the smaller number of latent variable locations where examinees are placed in DCMs. We demonstrate this result by comparing DCM and IRT model reliability for a series of models estimated with data from an end-of-grade test, leading to a discussion of how DCMs can be used to change the process character of large scale testing to precisely measure latent skills of examinees with fewer items or measure more dimensions with the same number of items. [Authors' abstract]

Templin, J., Roussos, L., & Stout, W. (2004, March). *Modeling ordered polytomous attributes through ordered dichotomous attributes*. Paper presented at Educational Testing Service, Princeton, New Jersey.

von Davier, M., DiBello, L., & Yamamoto, K. (2006). *Reporting test outcomes using models for cognitive diagnosis* (Research Report RR-06-28). Princeton, NJ: Educational Testing Service.

Models for cognitive diagnosis have been developed as an attempt to provide more than a single test score from item response data. Most approaches are based on a hypothesis that relates items to underlying skills. This relation takes the form of a design matrix that specifies for each cognitive item which skills are required to solve the item and which are not. This report outlines one direction that developments of cognitive diagnosis models are taking. It does not claim completeness, but describes a line of models that can be traced back to Tatsuoka's seminal work on the rule space methodology and that finds its current form in models that combine features of confirmatory latent factor analysis, multiple classification latent class models, and multidimensional item response models. [Authors' abstract]

Yan, D., Almond, R., & Mislevy, R. (2003, April). *Empirical comparisons of cognitive diagnostic models*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago.

Yen, W. M. (1987, June). *A Bayesian/IRT index of objective performance*. Paper presented at the meeting of the Psychometric Society, Montreal, Quebec, Canada.

Zhou J., Gierl, M., & Cui, Y. (2009, April). *Attribute reliability in cognitive diagnostic assessment*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Market basket reporting

Colvin, R. L. (2000, February). *NAEP market-basket reporting: A journalist's perspective*. Paper presented at the workshop of the Committee on NAEP Reporting Practices: Investigating District-Level and Market-Based Reporting, National Research Council, Washington, DC.

DeVito, P. J., & Koenig, J. A. (Eds.). (2000). *Designing a market-basket for NAEP: Summary of a workshop*. Washington, DC: National Academy Press. Retrieved March 31, 2009, from http://www.nap.edu/catalog.php?record_id=9891

Educational Testing Service. (1998). Prepare for mathematics market basket (Chapter 11) and analyze and report on mathematics market basket booklet (Chapter 18, Task 52). In *NAEP 2000: Application for cooperative agreement for the National Assessment of Educational Progress—Technical application*. Author.

Kenney, P. A. (2000). *Market basket reporting for NAEP: A content perspective*. Paper presented at the March workshop of the Committee on NAEP Reporting Practices: Investigating District-Level and Market-Based Reporting, National Research Council, Washington, DC.

Kolstad, A. (2000, February). *Simplifying the interpretation of NAEP results with market baskets and shortened forms of NAEP*. Paper presented at the workshop of the Committee on NAEP Reporting Practices: Investigating District-Level and Market-Based Reporting, National Research Council, Washington, DC.

Mazzeo, J. (2000, February). *NAEP's year-2000 market-basket study: What do we expect to learn?* Paper presented at the workshop of the Committee on NAEP Reporting Practices: Investigating District-Level and Market-Based Reporting, National Research Council, Washington, DC.

Mazzeo, J., Kulick, E., Tay-Lim, B., & Perie, M. (2006). *Technical report for the 2000 market-basket study in mathematics* (Research Report ETS-NAEP-06-T01). Princeton, NJ: Educational Testing Service.

This technical report presents the goals and design of the 2000 National Assessment of Educational Progress (NAEP) market-basket study, describes the analyses that were conducted to produce the prototype NAEP market-basket report card, and presents and discusses results from the study that are pertinent to selected technical and psychometric issues associated with the potential implementation of a market-basket reporting option for NAEP. A market basket is a specific collection of test items intended to be representative or illustrative of a domain of material included in an assessment. Reporting assessment results in terms of the scores on this collection of items and publicly releasing the items are what is typically meant by market-basket reporting. Two market-basket test forms were constructed and administered to nationally representative samples of fourth

grade students. Results for a nationally representative sample of students from both sets of projections were compared with each other and with the results actually obtained by directly administering the market basket to separate nationally representative samples. While the two kinds of projection results were generally similar, differences between them, consistent with what one would expect from basic measurement theory, were evident. Furthermore, both sets of projection results were similar, in most cases, to actual results obtained by directly administering the market baskets to separate, randomly equivalent samples. There were, however, some notable differences. [Authors' abstract]

McConachie, M. (2000, February). *State policy perspectives on NAEP market basket reporting*. Paper presented at the workshop of the Committee on NAEP Reporting Practices: Investigating District-Level and Market-Based Reporting, National Research Council, Washington, DC.

Mislevy, R. J. (1998). Implications of market-basket-reporting for achievement level setting. *Applied Measurement in Education, 11*(1), 49-63.

Discusses ways in which reporting National Assessment of Educational Progress (NAEP) results in terms of a market basket of tasks would affect achievement-level reporting. After reviewing current NAEP reporting and achievement-level setting procedures, 3 market-basket variations are described. Ways in which achievement-level standards would be set, interpreted, and validated are then discussed. The conclusions are as follows: (a) the structure of the market-basket reporting scale can be exploited to simplify a key step in the standard-setting process, namely mapping item- or booklet-level judgments to the reporting scale; (b) the more transparent meaning of market-basket scores, in contrast to scaled scores and behavioral descriptions, clarifies the limitations of NAEP performances as evidence about the range of student proficiencies and accomplishments that the public's and educators' interests may span; and (c) market-basket reporting approaches that enable individual students to take a full market-basket set of items simplify data-gathering and analysis for validity studies of achievement-level set-points and interpretations. [Author's abstract]

National Assessment Governing Board. (1997). *Resolution on market basket reporting, report of August 2*. Washington, DC: Author.

Truby, R. (2000, February). *A market basket for NAEP: Policies and objectives of the National Assessment Governing Board*. Paper presented at the workshop of the Committee on NAEP Reporting Practices: Investigating District-Level and Market-Based Reporting, National Research Council, Washington, DC.

Reporting and validity

Brown, G., & Hattie, J. (2009, April). *Understanding teachers' thinking about assessment: Insights for developing better educational assessments*. Paper

presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

The studies of Assessment Tools for Teaching and Learning (asTTle) use have shown that how assessment is conceived and the beliefs that teachers have about assessment are associated with gains in student learning as well as more effective use of test reports. Hence, we suggest that the New Zealand example demonstrates that if test development takes into account the pre-existing conceptions of teachers about assessment, it will result in test reporting and professional development that are more effective in raising student achievement. This is so because teachers will be able to use the tests for improvement, while satisfying accountability-oriented requirements. Taking into account both of these purposes for assessment and devising an integrated reporting system that addresses them appropriately is an essential aspect of assessment *for* and *of* learning. [Authors' conclusion]

Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational Measurement: Issues and Practice*, 10(3), 3-9.

The scales of the National Assessment of Educational Progress (NAEP), as constructed, do not yield meaningful criterion-referenced interpretations. Poorly defined NAEP goals and the present knowledge base do not allow the measurement of what examinees can and cannot do. Inappropriate interpretations of NAEP data are discussed, with specific examples. [Author's abstract]

Gardner, E. (1989). *Five common misuses of tests*. ERIC Digest.

Five of the common misuses of tests are reviewed: (1) acceptance of the test title as an accurate and complete description of the variable being measured (failure to examine the manual and the items carefully to know the specific aspects to be tested can result in misuse through selection of an inappropriate test for a particular purpose or situation); (2) ignoring the error of measurement in test scores; (3) use of a single test score for decision making (scores are not interpreted in the full context of the various elements that characterize students, teachers, and the environment); (4) a lack of understanding of the meaning of test score reporting (the misinterpretation of raw scores or grade equivalents is common); and (5) attributing cause of behavior measured to test (confusing the information provided by a test score with interpretations of what caused the behavior or described by the score). A test score gives no information as to why the individual performed as reported. No statistical manipulation of test data will permit more than probabilistic inferences about causation or future performance. [Author's abstract]

Haertel, E. H. (1991). Reasonable inferences for the trial state NAEP given the current design: Inferences that can and cannot be made. In R. Glaser, R. Linn, & G.

Bohrnstedt (Eds.), *Assessing student achievement in the states: Background studies*. Stanford, CA: National Academy of Education.

Hattie, J. (2009, April). *Visibly learning from reports: The validity of score reports*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

This paper outlines a fundamental claim about the validity of Reports, and then via a series of empirical studies introduces a series of principles that aims to assist in maximizing the accuracy and appropriateness of interpretations of Reports. Two other sources of evidence are used to derive and defend additional principals - the human computer interface research and the findings from visual graphics. [Author's abstract]

Hattie, J. A. C., Brown, G. T. L., Keegan, P., Irving, E., & Mackay, A. (2005, June). *asTTle V4: Improving the planning and reporting of learning*. Paper presented to the NSADAP Conference, Auckland, New Zealand.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average". *Educational Measurement: Issues and Practice*, 9(3), 5-14.

Are all states and nearly all districts claiming that their students are above the national average? If so, are the test results "inflated and misleading?" What are the factors that contribute to the abundance of "above average" scores? [Authors' abstract]

Linn, R. L., & Hambleton, R. K. (1992). Customized tests and customized norms. *Applied Measurement in Education*, 4(3), 185-207.

Describes the four main approaches to customized educational testing. Ability of customized testing to yield both valid normative and curriculum-specific information; Threats to the validity of normative interpretations. [Authors' abstract]

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement: Issues and Practice*, 28(1), 3-9.

This article has three goals. The first goal is to clarify the role that the consequences of test score use play in validity judgments by reviewing the role that modern writers on validity have ascribed for consequences in supporting validity judgments. The second goal is to summarize current views on who is responsible for collecting evidence of test score use consequences by attempting to separate the responsibilities of the test developer and the test user. The last goal is to offer a framework that attempts to prescribe the conditions under which the

responsibility for collecting evidence of consequences falls to the test developer or to the test user. [Authors' abstract]

Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment, 13*(2), 108-131.

In the United States, when English language learners (ELLs) are tested, they are usually tested in English and their limited English proficiency is a potential cause of construct-irrelevant variance. When such irrelevancies affect test *scores*, inaccurate *interpretations* of ELLs' knowledge, skills, and abilities may occur. In this article, we review validity issues relevant to the *educational assessment* of ELLs and discuss methods that can be used to evaluate the degree to which *interpretations* of their test *scores* are valid. Our discussion is organized using the five sources of validity evidence promulgated by the Standards for Educational and Psychological Testing. Technical details for some validation methods are provided. When evaluating the validity of a test for ELLs, the evaluation methods should be selected so that the evidence gathered specifically addresses appropriate test use. Such evaluations should be comprehensive and based on multiple sources of validity evidence. [Authors' abstract]

Watermann, R., & Klieme, E. (2002). Reporting results of large-scale assessment in psychologically and educationally meaningful terms: Construct validation and proficiency scaling in TIMSS. *European Journal of Psychological Assessment, 18*(3), 190-203.

In their function as a specific form of evaluation in the educational system, large-scale assessments are used to describe overall structures, salient features, and outcomes of educational processes. Whether this kind of evaluation is meaningful on the system level, and whether its results are likely to be of use for classroom practice, teacher training, and curriculum design is wholly dependent on the validity of the test instruments. The issues here are the validity of instruments with regard to the curricula of different countries, the underlying proficiency dimensions, and the appropriate behavior-oriented criteria for the interpretation of test *scores*. Using the TIMSS secondary school study as an illustrative example, the authors discuss methods for the validation of large-scale assessments and present results from the field of mathematics. Analyses of the cognitive demands of test items based on psychological conceptualizations of mathematical problem solving are combined with a behavior-oriented interpretation of different levels of a latent proficiency scale. Results show that proficiency scaling is a useful heuristic tool that can be used to integrate test theory, cognitive psychology, and didactics, and provide a meaningful way of interpreting the results of studies. [Authors' abstract]