Resources for Reporting Test Scores: A Bibliography for the Assessment Community

[*Prepared for* the National Council on Measurement in Education; Version: 5/6/09]

Nina Deng and Hanwook Yoo [1,2,3]

Center for Educational Assessment
University of Massachusetts Amherst

[1] This work was entirely collaborative in nature and the order of authors is alphabetical.
[2] The authors are grateful to the National Council on Measurement in Education for its support in the compilation of this document.
[3] The authors also thank Ronald K. Hambleton and April L. Zenisky of the University of Massachusetts for substantial input and guidance throughout the preparation of this bibliography.

Resources for Reporting Test Scores:
A Bibliography for the Assessment Community

[Version: 5/6/09]

Resources for Reporting Test Scores:
A Bibliography for the Assessment Community

Introduction

Score reporting is a rising challenge for many testing agencies today regardless of the audience for the reports (local, state, national, and even international) and regardless of test purpose (norm-referenced and criterion-referenced achievement, diagnosis, growth, or credentialing). For example, NCLB requirements have shone a bright spotlight on K-12 assessment practices in the United States over the past several years (involving millions of reports to parents), and international comparisons of performance are likewise of great interest (and receive considerable attention from policy makers, educators, and the public).

Both here and overseas, educational tests are increasingly being used for a variety of important purposes, and in the realm of professional credentialing test results are high-stakes for individuals (and their professional organizations). Across testing contexts, stakeholders including the examinees themselves want results presented to them in ways that are clear, concise, and relevant. At the same time, score reporting has historically been a bit of a postscript to the test development process and has not always been held to the same quality standards as the assessments themselves. Research findings, too, to guide the process of score report design are often lacking. For agencies charged with developing score reporting resources, the literature on reporting is diffuse at best: it draws not only from psychometrics but also graphic design, cognitive psychology, public policy, public relations, and increasingly, information technology.

Clearly, score reporting is a rapidly evolving topic that simply cannot be done 'on the fly', and this bibliography endeavors to bring together references on numerous aspects of score reporting together as a resource for people involved in the development of score reports and reporting materials. Here, we have searched the psychometric literature as well of that of related fields to identify journal articles, technical reports and documents, and conference papers that could be used by testing agencies to inform reporting practices in a variety of testing contexts.

We begin our bibliography by citing the relevant professional guidelines in the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999), the *Code of Professional Responsibilities in Educational Measurement* (NCME, 1995), and the *Code of Fair Testing Practices in Education* (*Code of Fair Testing Practices in Education*, 2004). From there we have organized the references into categories as noted in the Table of Contents:

- Guidelines references are those which offer readers general and specific guidance for report development in the form of principles or other advice.

- Report Levels and Audiences references are materials that concern reporting for different stakeholder groups and at different levels of aggregation (students, district, state, etc.)

- Scores and Reporting Contexts as a category encompasses a range of materials on the contents of score reports, including references on scale scores, achievement levels, scale anchoring/item mapping, domain score/subscore reporting, diagnostic score reporting, market basket reporting, and reporting and validity.

- Displaying Data and Accessing Results references address graphic design, report formatting, and reporting medium such as online reporting.

- Reporting Policy and Accountability references speak generally to the topic of reporting materials including reporting in an accountability context.

- Sample Reports provides references to a number of individual- and group-level score reports and interpretive guides that intended users of this bibliography may find useful as examples of current practices (please note that inclusion does not imply endorsement or formal review with respect to professional standards or other guidelines cited previously).

1. Professional Standards

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.

Relevant standards for score and test interpretations, score reports, and test uses follow:

Score Interpretation
1.1/ 1.2 / 1.3 / 1.4: Validity of score interpretation
2.11 / 7.3: Inferences with subpopulation
3.4: Documenting normative or standardization samples or the criterion
4.10 / 10.11 / 13.4: Score comparability (local norms)
4.19 / 4.20: Cut scores
5.10 / 11.17 / 11.18 / 13.12 / 13.13: Interpretive material for local release
6.5: Using statistical descriptions and analyses (raw and derived score & standard error)
11.15 / 13.15 / 15.12: Appropriate contextual information (potential misinterpretations)

Score Reporting
5.11 / 6.12 / 12.15: Give more information for computer-generated interpretation
7.5 / 11.20 / 12.19 / 13.7: Need of description and analysis of alternate explanations
7.2 / 7.8 / 13.19: For subgroups (gender, age, ethnicity, sample size & distribution)
8.8: Categorical decisions to assign individuals
8.9: Confidentiality to report scores
9.5 / 10.11: Do not report flagged scores
11.6 / 12.9 / 12.20 / 13.14 / 15.11: Format appropriate for recipient
13.16: Date of test administration and the age of any norms to interpret report
13.17 / 15.3 / 15.4: Definition of score & technical support need to use of gained score

Test Interpretation
2.2 / 2.3: Standard error of measurement to interpret individual score
12.14 / 12.16: Considering possible conditions for each examinee before interpret

Use of Test Scores
7.10 / 7.11: Mean test score differences between relevant subgroups (construct-irrelevant)

Code of Fair Testing Practices in Education. (2004). Washington, DC: Joint Committee on Testing Practices. Retrieved March 31, 2009, from http://www.apa.org/science/fairtestcode.html

For test developers: test developers should report test results accurately and provide information to help test users interpret test results correctly.
C-1. Provide information to support recommended interpretations of the results, including the nature of the content, norms or comparison groups, and other technical evidence. Advise test users of the benefits and limitations of test results and their interpretation. Warn against assigning greater precision than is warranted.
C-2. Provide guidance regarding the interpretations of results for tests administered with modifications. Inform test users of potential problems in interpreting test results when tests or test administration procedures are modified.
C-3. Specify appropriate uses of test results and warn test users of potential misuses.
C-4. When test developers set standards, provide the rationale, procedures, and evidence for setting performance standards or passing scores. Avoid using stigmatizing labels.
C-5. Encourage test users to base decisions about test takers on multiple sources of appropriate information, not on a single test score.
C-6. Provide information to enable test users to accurately interpret and report test results for groups of test takers, including information about who were and who were not included in the different groups being compared, and information about factors that might influence the interpretation of results.
C-7. Provide test results in a timely fashion and in a manner that is understood by the test taker.
C-8. Provide guidance to test users about how to monitor the extent to which the test is fulfilling its intended purposes.

For test users: test users should report and interpret test results accurately and clearly.
C-1. Interpret the meaning of the test results, taking into account the nature of the content, norms or comparison groups, other technical evidence, and benefits and limitations of test results.
C-2. Interpret test results from modified test or test administration procedures in view of the impact those modifications may have had on test results.
C-3. Avoid using tests for purposes other than those recommended by the test developer unless there is evidence to support the intended use or interpretation.
C-4. Review the procedures for setting performance standards or passing scores. Avoid using stigmatizing labels.
C-5. Avoid using a single test score as the sole determinant of decisions about test takers. Interpret test scores in conjunction with other information about individuals.
C-6. State the intended interpretation and use of test results for groups of test takers. Avoid grouping test results for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use. Report procedures that were followed in determining who were and who were not

included in the groups being compared and describe factors that might influence the interpretation of results.
C-7. Communicate test results in a timely fashion and in a manner that is understood by the test taker.
C-8. Develop and implement procedures for monitoring test use, including consistency with the intended purposes of the test.

Informing test takers: test developers or test users should inform test takers about the nature of the test, test taker rights and responsibilities, the appropriate use of scores, and procedures for resolving challenges to scores.
D–1. Inform test takers in advance of the test administration about the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. Make such information available to all test takers.
D–2. When a test is optional, provide test takers or their parents/guardians with information to help them judge whether a test should be taken—including indications of any consequences that may result from not taking the test (e.g., not being eligible to compete for a particular scholarship)—and whether there is an available alternative to the test.
D–3. Provide test takers or their parents/guardians with information about rights test takers may have to obtain copies of tests and completed answer sheets, to retake tests, to have tests rescored, or to have scores declared invalid.
D–4. Provide test takers or their parents/guardians with information about responsibilities test takers have, such as being aware of the intended purpose and uses of the test, performing at capacity, following directions, and not disclosing test items or interfering with other test takers.
D–5. Inform test takers or their parents/guardians how long scores will be kept on file and indicate to whom, under what circumstances, and in what manner test scores and related information will or will not be released. Protect test scores from unauthorized release and access.
D–6. Describe procedures for investigating and resolving circumstances that might result in canceling or withholding scores, such as failure to adhere to specified testing procedures.
D–7. Describe procedures that test takers, parents/guardians, and other interested parties may use to obtain more information about the test, register complaints, and have problems resolved.

National Council on Measurement in Education. (1995). *Code of professional responsibilities in educational measurement.* Washington, DC: Author. Retrieved March 31, 2009, from http://www.natd.org/Code_of_Professional_Responsibilities.html

6.1 Conduct these activities in an informed, objective, and fair manner within the context of the assessment's limitations and with an understanding of the potential consequences of use.

6.2 Provide to those who receive assessment results information about the assessment, its purposes, its limitations, and its uses necessary for the proper interpretation of the results.
6.3 Provide to those who receive score reports an understandable written description of all reported scores, including proper interpretations and likely misinterpretations.
6.4 Communicate to appropriate audiences the results of the assessment in an understandable and timely manner, including proper interpretations and likely misinterpretations.
6.5 Evaluate and communicate the adequacy and appropriateness of any norms or standards used in the interpretation of assessment results.
6.6 Inform parties involved in the assessment process how assessment results may affect them.
6.7 Use multiple sources and types of relevant information about persons or programs whenever possible in making educational decisions.
6.8 Avoid making, and actively discourage others from making, inaccurate reports, unsubstantiated claims, inappropriate interpretations, or otherwise false and misleading statements about assessment results.
6.9 Disclose to examinees and others whether and how long the results of the assessment will be kept on file, procedures for appeal and rescoring, rights examinees and others have to the assessment information, and how those rights may be exercised.
6.10 Report any apparent misuses of assessment information to those responsible for the assessment process.
6.11 Protect the rights to privacy of individuals and institutions involved in the assessment process.

2. Guidelines

Allalouf, A. (2007). An NCME instructional module on quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice, 26*(1), 36-46.

There is significant potential for error in long production processes that consist of sequential stages, each of which is heavily dependent on the previous stage, such as the SER (Scoring, Equating, and Reporting) process. Quality control procedures are required in order to monitor this process and to reduce the number of mistakes to a minimum. In the context of this module, quality control is a formal systematic process designed to ensure that expected quality standards are achieved during scoring, equating, and reporting of test scores. The module divides the SER process into 11 steps. For each step, possible mistakes that might occur are listed, followed by examples and quality control procedures for avoiding, detecting, or dealing with these mistakes. Most of the listed quality control procedures are also relevant for Internet-delivered and scored testing. Lessons from other industries are also discussed. The motto of this module is: There is a reason for every mistake. If you can identify the mistake, you can identify the reason it happened and prevent it from recurring. [Author's abstract]

Aschbacher, P. R., & Herman, J. L. (1991). *Guidelines for effective score reporting* (CSE Technical Report 326). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

The paper examines the practice in state reporting of assessment results based on 1984 and 1989 reviews from over 30 states, and to provide guidelines for effective reporting, derived from the literature on cognitive psychology, communication, and information representation and decision-making, along with illustrations of exemplary practice. Both content and format concerns are addressed. [Authors' abstract]

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145-220.

A critical, but often neglected, component of any large-scale assessment program is the reporting of test results. In the past decade, a body of evidence has been compiled that raises concerns over the ways in which these results are reported to and understood by their intended audiences. In this study, current approaches for reporting student-level results on large-scale assessment were investigated. Recent student test score reports and interpretive guides from 11 states, three U.S. commercial testing companies, and two Canadian provinces were reviewed. On the basis of past score-reporting research, testing standards, and the requirements of the *No Child Left Behind Act of 2001*, a number of promising and potentially problematic features of these reports and guides are identified, and

recommendations are offered to help enhance future score-reporting designs and to inform future research in this important area. [Authors' abstract]

Forte Fast, E., Blank, R. K., Potts, A., & Williams, A. (2002). *A guide to effective accountability reporting*. Washington, DC: Council of Chief State School Officers. Retrieved March 31, 2009, from http://www.ccsso.org/content/pdfs/GEAR.pdf

A Guide to Effective Accountability Reporting is intended to serve as a resource for the staffs of state education agencies (SEAs) and local education agencies (LEAs) who are responsible for producing state, district, or school report cards of the type required under many state or district accountability systems as well as under NCLB. This guide is not intended to provide an academic discussion of the nature of indicators and indicator systems, nor is it meant to cover the broad territory of accountability issues. It is meant to provide a resource for agencies, and to spur the thought of practitioners, as accountability reporting systems are tooled to meet the requirements of NCLB. [Authors' abstract]

Mills, C. N., & Hambleton, R. K. (1980, April). *Guidelines for reporting criterion-referenced test score information.* Paper presented at the meeting of the American Educational Research Association, Boston, MA.

General guidelines exist for reporting and interpreting test scores, but there are short comings in the available technology, especially when applied to criterion-referenced tests. Concerns that have been expressed in the educational measurement literature address the uses of test scores, the manner of reporting scores, limited testing knowledge among users, presentation of results to parents and students, and use of computer technology to report test scores. Several activities must occur before high quality test score reports can be prepared. These activities include the specification of information needs, building a testing program consistent with needs, identification of audiences and their levels of testing knowledge, proper test selection, and proper test construction. A rating system which can be used in designing or evaluating criterion-referenced test score reporting systems is presented, based on a logical analysis of criterion-referenced tests; current uses of the tests; and information needs of parents and students, building administrators, and district administrators. This rating system is organized around seven major categories of concern: information about objectives, information at the item level, information at the objective level, information at the subtest level, subject summaries, specialized services, and general services. [Authors' abstract]

National Education Goals Panel (NEGP). (1998). *Talking about tests: An idea book for state leaders.* Washington, DC: US Government Printing Office. Retrieved July 23, 2007, from http://govinfo.library.unt.edu/negp/REPORTS/98talking.PDF

Clear communication with parents about educational reform issues and the implementation of standards is essential. This publication presents ideas for state leaders on how better to inform parents about statewide assessments and how to report the results of these assessments to parents so that the results are more meaningful. The first section provides the perspectives of a parent and a policymaker when confronted with a new statewide test for the first time. The second section makes five strategic and four content recommendations and gives examples of how to make parents more aware of new tests, their purposes, and the changes they may bring. Section 3 gives some ideas on how to report testing results to parents. In the fourth section, five organizations that are committed to better communication with parents are described. Their structures, the coalitions they have built, and the products they produce to communicate with parents are described. Section 5 contains suggestions from the states with the best success in communicating with parents. This "Idea Book" also contains a series of "Close-ups" that provide stories from states on a variety of issues related to statewide testing, including reporting scores, evaluating communication tools, helping teachers with communication, and negotiating with the test provider. The appendices contain some annotated score reports, a set of resources to assist states in communicating with parents, and acknowledgments. [Author's abstract]

Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 677-710). Mahwah, NJ: Lawrence Erlbaum Associates.

Ysseldyke, J., & Nelson, J. R. (2002). Reporting results of student performance on large-scale assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 467-480). Mahwah, NJ: Lawrence Erlbaum Associates.

Identifies the characteristics of good state assessment and accountability reports on the scores of student performance on large-scale assessments, including the performance of students with disabilities. First, the authors consider what state and district reports should look like with specific consideration to issues of content. The authors describe ways in which these reports should be formatted and review the research on what the reports actually look like. A brief section is included on the actual results that state report on the performance, participation, and progress of students with disabilities. It is argued that reports should be clear, comprehensive, comparative, concise, and include confidentiality and cautionary statements. The authors also stress that the reports should be readable, responsive to audience needs, and well-organized. The chapter concludes by raising cautions about factors that lead to misinterpretation of data on trends in gaps between the performance of students with and without disabilities. [Authors' abstract]

3. Report Levels and Audiences

A-Plus Communications. (1999). *Reporting results: What the public wants to know*. A companion piece to 1999 issue of Education Week's "Quality Counts." Arlington, VA: Author.

Beaton, A. E. (1992). *Methodological issues in reporting NAEP results at district and school levels*. Paper commissioned by the National Assessment Governing Board.

Berends, M., & Koretz, D. M. (1995). Reporting minority students' test scores: How well can the National Assessment of Educational Progress account for differences in social context? *Educational Assessment, 3*(3), 249-285.

This article investigates the adequacy of the National Assessment of Educational Progress (NAEP) for taking into account dissimilarities in students' family, school, and community contexts when reporting test score differences among population groups (i.e., racial and ethnic minorities). This question was addressed by comparing the NAEP to other representative data for Grades 8 and 12--the National Education Longitudinal Study (NELS) and High School and Beyond (HSB)--that contain richer social context measures. Our analyses show that NAEP lacks a number of important social context measures and that the quality of some (but by no means all) of NAEP's measures is low because of reliance on student self-reports and other unreliable data sources. These weaknesses of NAEP have important practical implications: Compared to HSB and NELS, NAEP usually overestimates the achievement differences between students who come from different population groups but similar social contexts. However, at the secondary school level at which these analyses were conducted, these overestimates primarily reflect NAEP's lack of important measures rather than its reliance on student self-reports. [Authors' abstract]

Breithaupt, K., & Chuah, D. (2009, April). *Performance reporting for a licensing exam: What can, and should, we tell test takers?* Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Bunch, M. B. (1986, April). *Building a user-oriented statewide score reporting system.* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Burstein, L. (1990). Looking behind the "average": How are states reporting test results? *Educational Measurement: Issues and Practice, 9*(3), 23-26.

Means of interpreting norm-referenced *tests* to lead to more accurate *reporting results* are discussed, with particular emphasis on state-level and district-level data. Suggestions fall into the categories of documentation, frequency norm, and multiple form use. [Author's abstract]

In 1983 the Maryland State Department of Public Education (MSDE) issued a request for proposals for "The Development of the Score Reporting System for the Maryland Functional Testing Program." The MSDE called for a literature review, a national survey, a statewide survey of user needs and capabilities, an assessment of the state's report producing capability, and a final design for reports and a user's manual. Following a literature search, national and statewide surveys of reporting practices and information needs were conducted by Measurement Incorporated. Common and unique needs of district and building administrators, teachers and counselors, and parents and students were found. Using the nationwide search results, the information needs of students, parents, teachers, guidance counselors, principals, and district administrators in Maryland were surveyed. Score report design was based upon these studies emphasizing the accountability function of the tests. Four levels of reporting and seven content areas necessitated 28 separate score reports. Examples of four levels of reports (student, class, school, and local education agency) are presented. Each report is oriented to a specific audience, visual clutter is reduced, and diagnostic information is briefly presented. A user's guide provides thorough background on score interpretation at multiple levels. This score reporting system appears to meet the responsibilities and information needs of all its audiences. [Author's abstract]

Cieslak, P. (2000, February). *Milwaukee's experience with district-level NAEP results*. 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.). (1999). *Reporting district-level NAEP data: Summary of a workshop*. Washington, DC: National Academy Press. Retrieved March 31, 2009, from http://www.nap.edu/catalog.php?record_id=9768

DeVito, P. J., & Koenig, J. A. (Eds.). (2001). *NAEP reporting practices: Investigating district-level and market-basket reporting.* Washington, DC: National Academy Press. Retrieved March 31, 2009, from http://www.nap.edu/catalog.php?record_id=10049

Study questions focused on the: characteristics and features of the reporting methods, information needs likely to be served, level of interest in the reporting practices, types of inferences that could be based on the reported data, implications of the reporting methods for NAEP, and implications of the reporting methods for state and local education programs. [Authors' abstract]

Hambleton, R. K. (2002). How can we make NAEP and state test score reporting scale and reports more understandable? In R. W. Lissitz & W. D. Schafer (Eds.), *Assessment in educational reform* (pp. 192-205). Boston, MA: Allyn & Bacon.

Hambleton, R. K. (2002, February). *A new challenge: Making results from large scale assessments understandable and useful.* An invited presentation at the Provincial

Testing in Canadian Schools: Research, Policy, and Practice Conference, Victoria, British Columbia.

Hambleton, R. K., & Meara, K. (2000). Newspaper coverage of NAEP results, 1990 to 1999. In National Assessment Governing Board (Ed.), *Student performance standards of the National Assessment of Educational Progress: Affirmation and improvements* . Washington, DC: Editor.

Hambleton, R. K., & Slater, S. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report 430). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Teaching. Retrieved March 31, 2009, from http://research.cse.ucla.edu/Reports/TECH430.pdf

This research study is a follow-up to several recent studies conducted on NAEP reports that found policy makers and the media were misinterpreting test, figures, and tables. Our purposes were (a) to investigate the extent to which NAEP Executive Summary Reports are understandable to policy makers and educators, and (b) to the extent that problems are identified. Several recommendations are offered for improving the NAEP reports: First, all displays of data should be field tested prior to their use in NAEP Executive Summary Reports. A second recommendation is that NAEP reports for policy makers and educators should be considerably simplified. A third recommendation is that NAEP reports tailored to particular audiences may be needed to improve clarity, understandability, and usefulness. [Authors' abstract]

Hambleton, R. K., & Smith, T. (1999). *A focus group study of the general/public 1996 NAEP Science Reports* (Laboratory of Psychometric and Evaluative Research Report No. 361). Amherst, MA: University of Massachusetts, School of Education.

Haney, W., & Madaus, G. F. (1991). Caution on the future of NAEP: Arguments against using NAEP tests and data reporting below the state level. In R. Glaser, R. Linn, & G. Bohrnstedt (Eds.), *Assessing student achievement in the states: Background studies*. Stanford, CA: National Academy of Education.

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Teachers' ability to interpret standardized test scores. *Educational Measurement: Issues and Practice, 10*(4), 16-18.

To what extent do teachers possess the competence to interpret state testing program results properly? [Authors' abstract]

Jaeger, R. M. (1996). *Reporting large scale assessment results for public consumption: Some propositions and palliatives*. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.

Johnson, E. G. (1994). *Standard errors for below-state reporting of National Assessment of Educational Progress*. Paper prepared for the National Assessment Governing Board. Princeton, NJ: Educational Testing Service.

Koretz, D. M. (1991). State comparisons using NAEP: Large costs, disappointing benefits. *Educational Researcher*, *20*(3), 19-21.

Suggests that the proposed state-by-state National Assessment of Educational Progress (NAEP) will be unable to provide information about which state programs are responsible for differences in test scores. Raises concerns about its cost effectiveness and potential loss of validity if used in state comparisons. [Author's abstract]

Koretz, D., & Diebert, E. (1993). *Interpretations of National Assessment of Educational Progress (NAEP) anchor points and achievement levels by the print media in 1991.* Santa Monica, CA: RAND.

Levine, R., Rathbun, A., Selden, R., & Davis, A. (1998). *NAEP's constituents: What do they want? Report of the National Assessment of Educational Progress Constituents Survey and Focus Groups* (NCES 98–521). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

McDonnell, L. M. (1994). *Policymakers' views of student assessment*. Report commissioned by the Office of Educational Research and Improvement, U.S. Department of Education. Santa Monica, CA: RAND Institute on Education and Training.

O'Reilly, J. (2000, February). *District level and market-basket reporting: A district 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.

Patelis, T., & Matos, H. (2009, April). *Efforts to produce relevant score reports to school, district, and state officials on national tests.* Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

A historical overview of score reporting at the College Board is documented within this paper. Efforts to make score reports more meaningful and valuable to score reports users are described through the developmental activities that were underway during the production of the College Board's SAT Skills Insight reports for both students and state officials. Reflections of lessons learned throughout the report development process are also provided, along with the College Board's vision for future score reports. [Authors' abstract]

Robert, E. D. (1994, February). *Guidelines for the use of NAEP at the district and school levels*. Paper commissioned by the National Assessment Governing Board.

Rust, K. (1999). *NAEP sample designs and district level reporting*. Paper prepared for the National Research Council Workshop on District-Level Reporting, Washington, DC.

Selden, R. (1991). The case for district- and school-level results from NAEP. In R. Glaser, R. Linn, & G. Bohrnstedt (Eds.), *Assessing student achievement in the states: Background studies*. Stanford, CA: National Academy of Education.

Sicoly, F. (2002). What do school-level scores from large-scale assessments really measure? *Educational Measurement: Issues and Practice, 21*(4), 17-26.

Although assessments of mathematics, reading, and writing are assumed to measure distinct academic skills, this may be difficult owing to the pervasive influence of general ability on performance. Factor analyses of school-level data from 14 large-scale assessment programs revealed that 80% of the variance in mathematics, reading, and writing scores was due to a common, underlying factor. Multiple regression analyses confirmed that scores contribute little information that is unique to a particular subject (6% or less). Although different assessments may create the illusion of providing unique information, they may be tapping into generic cognitive abilities that cut across content areas. These results raise suspicions about the value and validity of interpretations based on school-level subject area scores. [Author's abstract]

Simmons, C., & Mwalimu, M. (2000). What NAEP's publics have to say. In M. L. Bourque & S. Byrd (Eds.), *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements. A study initiated to examine a decade of achievement level setting on NAEP* (pp. 184-219). Washington, DC: National Assessment Governing Board.

Sopko, K., & Reder, N. (2007). *Public and parent reporting requirements: NCLB and IDEA regulations*. In Forum. Alexandria, VA: National Association of State Directors of Special Education.

Trout, D. L., & Hyde, B. (2006, April). *Developing score reports for statewide assessments that are valued and used: Feedback from K-12 stakeholders.* Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Westin, T. (1999). *Reporting issues and strategies for disabled students in large scale assessments*. Washington, DC: Assessing Special Education Students, SCASS, CCSSO.

Ysseldyke, J., & Bielinski, J. (2002). Effect of different methods of reporting and reclassification on trends in test scores for students with disabilities. *Exceptional Children, 68*(2), 189-200.

State education agencies are now required to report on the educational performance and progress of all students, including students with disabilities. States are beginning to report trends, and to compare trends in performance of students with and without disabilities. We compare the effects of different methods of analyzing trends to illustrate how failure to account for changes in classification of students will lead to misinterpretation of data on the performance and progress of students with disabilities, and inappropriate policy decisions. We compare three ways of looking at trends over time, and use data from 5 years of assessment in a large state to illustrate the effects of students who change classification. We discuss how accounting for changes in classification of individual students will lead to more appropriate decisions and help avoid negative consequences for students with disabilities. [Authors' abstract]

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 sec- tion 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 then 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 methods 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 subscoring.* 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 subscoring 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 subscoring 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]

McPeek, 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 axe 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. Brennen (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 narket-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-1evel 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]

5. Displaying Data and Accessing Results

Bennett, K. B., & Flach, J. M. (1992). Graphical display: Implications for divided attention, focused attention, and problem-solving. *Human Factors, 34*(5), 513-533.

> When completing tasks in complex, dynamic domains observers must consider the relationships among many variables (e.g., integrated tasks) as well as the values of individual variables (e.g., focused tasks). A critical issue in display design is whether or not a single display format can achieve the dual design goals of supporting performance at both types of tasks. We consider this issue from a variety of perspectives. One relevant perspective is the basic research on attention and object perception, which concentrates on the interaction between visual features and processing capabilities. The principles of configurality are discussed, with the conclusion that they support the possibility of achieving the dual design goals. These considerations are necessary but not sufficient for effective display design. Graphic displays map information from a domain into visual features; the tasks to be completed are defined in terms of the domain, not in terms of the visual features alone. The implications of this subtle but extremely important difference are discussed. The laboratory research investigating alternative display formats is reviewed. Much like the attention literature, the results do not rule out the possibility that the dual design goals can be achieved. [Authors' abstract]

Best, L. A., Smith, L. D., & Stubbs, D. A. (2001). Graph use in psychology and other sciences. *Behavioural Processes, 54*(3), 155-165.

> Since the early 19th century, graphs have been recognised as an effective method of analysing and representing scientific data. However, levels of graph use have varied widely since then, partly due to increasing reliance on inferential statistics in some fields. Recent studies indicate that graph use is closely related to the 'hardness' of scientific disciplines, and that this finding holds for journal articles and textbooks across the subfields of psychology. In the area of animal behaviour, journals devote about one-sixth of their page space to graphs, a level of graph use approximating that of biology and physics. Implications for the training of scientists in the use of visual displays are considered. [Authors' abstract]

Carswell, C. M., Frankenberger, S., & Bernhard, D. (1991). Graphing in depth: Perspectives on the use of three-dimensional graphs to represent lower-dimensional data. *Behaviour and Information Technology*, *10*(6), 459-474.

Carswell, C. M., & Ramzy, C. (1997). Graphing small data sets: Should we bother? *Behaviour and Information Technology*, *16*(2), 61-71.

> While display designers tend to agree that the communication of large amounts of quantitative information calls for the use of graphs, there is less consensus about whether graphs should be used for small, summarized data sets. In the present study, three groups of 16 subjects viewed 11 sets of time series data presented as

tables, bar charts, or line graphs. Data sets varied in size (4, 7, or 13 values) and complexity (number and type of departures from linearity). Subjects provided written interpretations of each of the data sets, and these interpretations were scored for (1) overall number of propositions pertaining to the data set as a whole (global content), (2) number of propositions describing relations within a subset of the data (local content), and (3) number of references to specific data values (numeric content). For the larger (7- and 13-point) data sets, interpretations based on bar charts included the greatest overall global content, but line graph interpretations proved to be most sensitive to the actual information content (complexity) of the data sets. The greater sensitivity of the line graphs was still obtained with four-point data sets; however, this advantage was greater for men than for women. For data sets of all sizes, but especially for the smallest sets, gender differences in interpretation content were obtained. These differences are discussed within the context of more general individual differences presumed to exist in graph-reading strategies. [Authors' abstract]

Cleveland, W. S. (1985). *The elements of graphing data*. Monterey, CA: Wadsworth.

Reviewed by Simon, G. (1987). The elements of graphing data (book). *Journal of the American Statistical Association, 82*(397), 348-349.

Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.

Reviewed by Welsh, A. H. (1994). Visualizing data (book). *Journal of the American Statistical Association, 89*(427), 1136-1138.

Cleveland, W. S. (1994). *The elements of graphing data*. Summit, NJ: Hobart Press.

Reviewed by Ziegel, E. R. (1997). Book reviews. *Technometrics, 39*(2), 237-238.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphic methods. *Journal of the American Statistical Association*, *79*(387), 531-534.

The subject of graphical methods for data analysis and for data presentation needs a scientific foundation. In this article we take a few steps in the direction of establishing such a foundation. Our approach is based on graphical perception-the visual decoding of information encoded on graphs-and it includes both theory and experimentation to test the theory. The theory deals with a small but important piece of the whole process of graphical perception. The first part is an identification of a set of elementary perceptual tasks that are carried out when people extract quantitative information from graphs. The second part is an ordering of the tasks on the basis of how accurately people perform them. Elements of the theory are tested by experimentation in which subjects record their judgments of the quantitative information on graphs. The experiments validate these elements but also suggest that the set of elementary tasks should be

expanded. The theory provides a guideline for graph construction: Graphs should employ elementary tasks as high in the ordering as possible. This principle is applied to a variety of graphs, including bar charts, divided bar charts, pie charts, and statistical maps with shading. The conclusion is that radical surgery on these popular graphs is needed, and as replacements we offer alternative graphical forms-dot charts, dot charts with grouping, and framed-rectangle charts. [Authors' abstract]

Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, *229*, 828-833.

Graphical perception is the visual decoding of the quantitative and qualitative information encoded on graphs. Some recent theoretical/experimental investigations of graphical perception are described, identifying certain elementary graphical-perception tasks that are performed in the visual decoding of quantitative information from graphs. [Authors' abstract]

Dent, B. D. (1999). *Cartography: Thematic map design* (5th ed.). New York, NY: McGraw Hill.

Reviewed by Macdonald, A., & Mackaness, W. A. (2000). Book reviews. *International Journal of Geographical Information Science, 14*(4), 407-409.

Gillian, D. J., Wickens, C. D., Hollands, J. G., & Carswell, C. M. (1998). Guidelines for presenting quantitative data in HFES publications. *Human Factors*, *40*(1), 28-41.

This article provides guidelines for presenting quantitative data in papers for publication. The article begins with a reader-centered design philosophy that distills the maxim "know thy user" into three components: (a) know your users' tasks, (b) know the operations supported by your displays, and (c) match user's operations to the ones supported by your display. Next, factors affecting the decision to present data in text, tables, or graphs are described: the amount of data, the readers' informational needs, and the value of visualizing the data. The remainder of the article outlines the design decisions required once an author has selected graphs as the data presentation medium. Decisions about the type of graph depend on the readers' experience and informational needs as well as characteristics of the independent (predictor) variables and the dependent (criterion) variable. Finally, specific guidelines for the design of graphs are presented. The guidelines were derived from empirical studies, analyses of graph readers' tasks, and practice-based design guidelines. The guidelines focus on matching the specific sensory, perceptual, and cognitive operations required to read a graph to the operations that the graph supports. [Authors' abstract]

Gilmore, A. & Hattie, J.A. (2001). Understanding usage of an internet based information resource for teachers: The Assessment Resource Banks. *New Zealand Journal of Educational Studies, 32*(2), 237-258.

Guerard, E. B. (2000, August 7). Web site lets parents compare their kids' test scores with peers'. *eSchool News*. Retrieved March 31, 2009, from http://www.eschoolnews.com/news/showstory.cfm?ArticleID=1337

Harris, R. L. (1997). *Information graphics: A comprehensive illustrated reference.* Mumbai, India: Jaico.

Reviewed by Wilson, R. D. (1998). Information graphics: A comprehensive illustrated reference. *Journal of the American Society for Information Science, 49*(4), 383-384.

Jacoby, W. G. (1997). *Statistical graphics for univariate and bivariate data.* Thousand Oaks, CA: Sage Publications. (Monograph #117)

The purpose of this monograph is to present the major techniques that fall under the general heading of statistical graphics used in the social sciences field. The primary focus of the discussion is on analytic graphics. In other words, I concentrate on graphical techniques that the researcher would employ as an integral part of the data analysis process. There is little explicit coverage of so-called presentational graphics or the kinds of displays that are intended primarily for communicating completed analyses to a lay audience. [Author's abstract]

Jacoby, W. G. (1998). *Statistical graphics for visualizing multivariate data.* Thousand Oaks, CA: Sage Publications. (Monograph #120)

This monograph will examine graphical displays that are useful for visualizing multivariate data. As such, it will pick up the discussion that was begun in the companion volume within this series, Statistical Graphics for Visualizing Univariate and Bivariate Data (W. G. Jacoby, 1997). The basic objective here is to obtain pictorial representations of quantitative information. Multivariate data pose special challenges for statistical graphics, beyond those encountered with univariate or bivariate data. The central problem is to represent information that can vary along several dimensions (typically, one for each variable) in a display medium that is almost always inherently 2-dimensional in nature--a printed page or computer display. [Author's abstract]

Knupp, T., & Ansley, T. (2008, March). *Online, state-specific assessment score reports and interpretive guides*. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.

The study was to (a) identify states with score reports and interpretive guides available via the Internet, (b) identify the characteristics of online test score information that meet score reporting standards as specified by government requirements and measurement experts, and (c) describe the utility of the online assessment score information. It is found that states reported their scores online

in a variety of ways; different score information was available, different file types were available for download, different grades had scores available, and the data were disaggregated relative to different groups of students. The interpretive guides were equally as variable. The materials most commonly found in interpretive guides included listing additional resources, giving the meaning of the scores, and stating the purpose of the assessment. Information about test score precision and common misinterpretations of test scores were least likely to be mentioned in the guides. [Authors' abstract]

Kosslyn, S. (1985). Graphics and human information processing. *Journal of the American Statistical Association*, *80*(391), 499-512.

Kosslyn, S. (1994). *Elements of graph design*. New York, NY: W. H. Freeman.

Reviewed by Schreiner, D. E., & Murphy, A. J. (1996). Book reviews. *Technical Communication, 43*(3), 286-289.

Krug, S. (2000). *Don't make me think! A common sense approach to web usability.* Indianapolis, IN: QUE.

Meagher-Lundberg, P. (2000). *Comparison variables useful to teachers in analysing assessment results* (Tech. Rep. No. 1). Auckland, NZ: University of Auckland, Project asTTle.

Meagher-Lundberg, P. (2001). *Output reporting design: Focus group 2* (Tech. Rep. No. 10). Auckland, New Zealand: University of Auckland.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97.

A variety of researches are examined from the standpoint of information theory. It is shown that the unaided observer is severely limited in terms of the amount of information he can receive, process, and remember. However, it is shown that by the use of various techniques, e.g., use of several stimulus dimensions, recoding, and various mnemonic devices, this informational bottleneck can be broken. [Author's abstract]

Milroy, R., & Poulton, E. C. (1978). Labeling graphs for increased reading speed. *Ergonomics*, *21*(1), 55-61.

Three methods of labeling graphs were compared:
1. direct labeling on the functions
2. a key inserted on the graph field below the functions
3. a key inserted below the figure in the position of the figure caption.
In both a separate-groups comparison and in a subsequent within-subjects comparison, direct labeling gave reliably the quickest readings ($p < 0.01$) without

loss of accuracy. Reading the labels directly appeared to involve fewer steps and depend less upon short-term memory. [Authors' abstract]

Monmonier, M. (1991). *How to lie with maps*. Chicago, IL: University of Chicago Press.

Reviewed by Morrison, P. (1991). Necessary white lies. *Scientific American, 265*(1), 124-125.

Nielsen, J. (1993). *Usability engineering*. San Diego, CA: Morgan Kaufmann.

Reviewed by Albers, M. J., & Lisberg, B. C. (2000). Information design: A bibliography. *Technical Communication, 47*(2), 170

Nielsen, J. (2000). *Design web usability: The practice of simplicity.* Indianapolis, IN: New Riders Publishing.

Norman, D. A. (2002). *The design of everyday things.* New York, NY: Basic Books.

Originally published as *The psychology of everyday things* in 1988 and reviewed by Watts, P. (1989). The psychology of everyday things (book review). *Management Review, 78*(5), 60-61.

Also reviewed by Schmeil, A. (2008). The design of everyday things (book review). *Studies in Communication Sciences, 8*(2), 408-410.

Norman, D. A. (2004). *Emotional design: Why we love (or hate) everyday things.* New York, NY: Basic Books.

Reviewed by Gold, S. F., Chenoweth, E., & Zaleski, J. (2003). Emotional Design: Why we love to hate everyday things (book). *Publishers Weekly, 250*(45), 50-51.

Pickle, L. W., & Herrmann, D. (1994). The process of reading statistical maps: The effect of color. *Statistical Computing and Statistical Graphics Newsletter*, *5*(1), 12-16.

Rubin, J. (1994). *Handbook of usability testing*. New York, NY: John Wiley & Sons.

Reviewed by Shaw, D. (1996). Handbook of usability testing: How to plan, design, and conduct effective tests. *Journal of the American Society for Information Science, 47*(3), 258-259.

Salvagno, M., & Teglasi, H. (1987). Teacher perceptions of different types of information in psychological reports. *Journal of School Psychology, 25*(4)*,* 4l5-424.

One hundred and sixty elementary school teachers rated the helpfulness of various types of information on test-based and observation-based reports. There was no difference between the test-based and the observation-based reports in overall

rating of helpfulness. On both types of reports, interpretive material was consistently rated as more helpful than factual or descriptive information in all content areas. On the test-based report, information about personality dynamics was rated as more helpful than description of intellectual functioning or academic achievement. Teachers desire interpretations that go beyond that which is immediately observable in the behavior or test data. They prefer recommendations that provide specific guidelines for implementation and are easy to carry out. The type of report as well as the gender of the child described in the report influenced teachers' responses to recommendations. [Authors' abstract]

Simkin, D., & Hastie, R. (1987). An information processing analysis of graph perception. *Journal of the American Statistical Association*, *82*(398), 454-465.

Recent work on graph perception has focused on the nature of the processes that operate when people decode the information represented in graphs. We began our investigations by gathering evidence that people have generic expectations about what types of information will be the major messages in various types of graphs. These graph schemata suggested how graph type and judgment type would interact to determine the speed and accuracy of quantitative information extraction. These predictions were confirmed by the finding that a comparison judgment was most accurate when the judgment required assessing position along a common scale (simple bar chart), had intermediate accuracy on length judgments (divided bar chart), and was least accurate when assessing angles (pie chart). In contrast, when the judgment was an estimate of the proportion of the whole, angle assessments (pie chart) were as accurate as position (simple bar chart) and more accurate than length (divided bar chart). Proposals for elementary information processes involving anchoring, scanning, projection, superimposition, and detection operators were made to explain this interaction. [Authors' abstract]

Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

The classic book on statistical graphics, charts, tables. Theory and practice in the design of data graphics, 250 illustrations of the best (and a few of the worst) statistical graphics, with detailed analysis of how to display data for precise, effective, quick analysis, design of the high-resolution displays, small multiples, editing and improving graphics, the data-ink ratio. Time-series, relational graphics, data maps, multivariate designs. Detection of graphical deception: design variation vs. data variation, sources of deceptions, esthetics and data graphical displays. [Copied March 31, 2009, from http://www.edwardtufte.com/tufte/books_vdqi]

Reviewed by Fienberg, S. E. (1985). The visual display of quantitative information (book review). *Journal of Economic Literature, 23*(4), 1798.

Tufte, E. R. (1990). *Envisioning information.* Cheshire, CT: Graphics Press.

This book celebrates escapes from the flatlands of both paper and computer screen, showing superb displays of high-dimensional complex data. The most design-oriented of Edward Tufte's books, Envisioning Information shows maps, charts, scientific presentations, diagrams, computer interfaces, statistical graphics and tables, stereo photographs, guidebooks, courtroom exhibits, timetables, use of color, a pop-up, and many other wonderful displays of information. The book provides practical advice about how to explain complex material by visual means, with extraordinary examples to illustrate the fundamental principles of information displays. Topics include escaping flatland, color and information, micro/macro designs, layering and separation, small multiples, and narratives. [Copied March 31, 2009, from http://www.edwardtufte.com/tufte/books_ei]

Reviewed by Morrison, P. (1990). Envisioning information (book). *Scientific American, 263*(4), 131.

Tukey, J. W. (1990). Data-based graphics: Visual display in the decades to come. *Statistical Science, 5*(3), 327-339. Retrieved March 31, 2009, from http://www.jstor.org/stable/2245820

Tversky, B., & Schiano, D. J. (1989). Perceptual and conceptual factors in distortions in memory for graphs and maps. *Journal of Experimental Psychology: General*, *118*(4), 387-398.

We propose that representations of visual stimuli are a consequence of both perceptual and conceptual factors that may be revealed in systematic errors in memory. Three experiments demonstrated increased (horizontal or vertical) symmetry in perception and memory of nearly symmetric curves in graphs and rivers in maps. Next, a conceptual factor, an accompanying description biasing toward symmetry or asymmetry, also distorted memory in the expected direction for the symmetric descriptions. In the two final experiments, we investigated conceptual factors in selection of a frame of reference. Subjects remembered lines in graphs, but not in maps, as closer to the imaginary 45° line. Combined with earlier research, this suggests that the reference frame for map lines is the canonical axes and for graph lines, the imaginary 45° line. [Authors' abstract]

U.S. Department of Health & Human Services. (2009). *Usability guide*. Retrieved March 31, 2009, from http://usability.gov/

Vernon, M. D. (1952). The use and value of graphical methods of presenting quantitative data. *Occupational Psychology*, *26*, 22-24.

Wainer, H. (1984). How to display data badly. *The American Statistician, 38*(1), 137-147.

The aim of good data graphics is to display data accurately and clearly. This definition is used as a point of departure for developing 12 rules of bad data

display: (1) show as little data as possible (minimize the data density); (2) hide what data you do show (minimize the data-ink ratio); (3) ignore the visual metaphor altogether; (4) only order matters (The Pravda School of Ordinal Graphics); (5) graph data out of context; (6) change scales in mid-axis; (7) emphasize the trivial (ignore the important); (8) jiggle the baseline; (9) order graphs and tables alphabetically; (10) label illegibly, incompletely, incorrectly, and ambiguously; (11) more is murkier (more decimal places and more dimensions); and (12) if it has been done well in the past, think of a new way to do it. Although the tone of this presentation is light, and points in the wrong direction, the aim is serious. The 12 "rules" point clearly toward an outlook that provides many hints for good display. [Author's abstract]

Wainer, H. (1990). Graphical visions from William Playfair to John Tukey. *Statistical Science, 5*(3), 340-346.

This paper discusses the similarities and differences in Playfair and Tukey's visions of what graphically displaying quantitative phenomena can do now, and might do in the future. As part of this discussion we examine: (1) how fundamental graphic tools have become to the scientist, (2) three instances where modern views of graphics are unchanged since Playfair's time, and (3) one area where there has been a change. The paper concludes with a discussion of five important areas of current and future graphic concern. [Author's abstract]

Wainer, H. (1990). Measuring graphicacy. *Chance, 3*(4), 52-58.

Wainer, H. (1991). Integrating figures and text. *Chance, 4*(3), 58-60.

Wainer, H. (1991). Elegance, grace, impact and graphical displays. *Chance, 4*(4), 45-47.

Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher, 21*(1), 14-23.

Quantitative phenomena can be displayed effectively in a variety of ways, but to do so, requires an understanding of both the structure of the phenomena and the limitations of candidate display formats. This article (a) recounts three historic instances of the vital role data displays played in important discoveries, (b) provides three levels of information that form the basis of a theory of display to help us better measure both display quality and human graphicacy, and (c) describes three steps to improve the quality of tabular presentation. [Author's abstract]

Wainer, H. (1993). Making readable overhead displays. *Chance, 6*(2), 46-49.

Wainer, H. (1993). Graphing multiple comparisons: Some comments on Tukey. *Journal of Computational and Graphical Statistics, 2*(1), 35-40.

Wainer, H. (1993). Graphical answers to scientific questions. *Chance, 6*(4), 48-50.

Wainer, H. (1996). Depicting error. *The American Statistician, 50*(2), 101-111.

Discusses the importance of minimizing errors in the presentation of data to prevent misinterpretation and incorrect inferences. Alternatives for the effective communication of errors; Tabular and graphical display of errors; Focus of the study on errors in means. [Author's abstract]

Wainer, H. (1996). Using trilinear plots for NAEP data. *Journal of Educational Measurement, 33*(1), 41-55.

Understanding the distribution of achievement levels of students' performance in the National Assessment of Educational Progress (NAEP) is aided through the use of the trilinear chart. In this article, this chart is described and its use illustrated with data from the 1992 state NAEP mathematics assessment. It is shown that one can see readily the trends in performance for different demographic groups for all of the 44 participating jurisdictions simultaneously. It is suggested that this graphical form may be useful in other contexts, as well. [Author's abstract]

Wainer, H. (1997). Improving tabular displays with NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics*, *22*(1), 1-30.

The modern world is rich with data; an inability to effectively utilize these data is a real handicap. One common mode of data communication is the printed data table. In this article we provide four guidelines the use of which can make tables more effective and evocative data displays. We use the National Assessment of Educational Progress both to provide inspiration for the development of these guidelines and to illustrate their operation. We also discuss a theoretical structure to aid in the development of test items to tap students' proficiency in extracting information from tables. [Author's abstract]

Wainer, H. (1997). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot*. New York, NY: Copernicus.

Reviewed by Ree, M. J., & Summers, L. (1998). Visual revelations: Graphic tales of fate and deception from napoleon bonaparte to ross perot. *Personnel Psychology, 51*(1), 226-229.

Wainer, H. (1997). Some multivariate displays in NAEP. *Psychological Methods, 2*(1), 34-63.

The principal goal of graphic display is to ease access to complex information. Simple univariate displays are easy to understand but usually do not have the capability to transmit accurately the often complex structure of multivariate data. Multivariate displays were specifically designed for exactly this purpose. The

National Assessment of Educational Progress (NAEP) generates data of a multivariate richness and complexity that defies accurate univariate transmission. The broad use and understanding of the information NAEP provides can be aided through the use of more suitable and evocative data displays. In this article, we demonstrate the limitations of univariate displays and suggest some multivariate displays that may enable us to understand, and thence communicate, what is contained in NAEP more fully. [Author's abstract]

Wainer, H. (2002). Clear thinking made visible: Redesigning score reports for students. *Visual Revelations, 15*(1), 56-58.

Wainer, H. (2009). *Picturing the uncertain world: How to understand, communicate, and control uncertainty through graphical display.* Princeton, NJ: Princeton University Press.

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, *36*(4), 301-335.

Presents a redesign and validity study of displays for communicating National Assessment of Educational Progress (NAEP) results in the United States. Use of the methodology to aid the evolution of data displays; Drawbacks of using the NAEP reports; Comparison between the accuracy of the original and redesigned formats. [Authors' abstract]

Wainer, H., & Thissen, D. (1981). Graphical data analysis. *Annual Review of Psychology, 32*, 191-241.

Focuses on the development of graphical methods for data analysis and communication. History of the use of visual displays to present quantitative materials; Characteristics of a graphical display; Usability of graphical displays. [Authors' abstract]

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2009, April). *At or above proficient: The reporting of NAEP results in the internet age.* Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

The purpose of this paper is to provide a brief overview of the web-based score reporting practices used by the National Assessment of Educational Progress (NAEP), as an example of the Internet as a score reporting (and test information) medium. For states and other groups involved in reporting the results of large-scale educational tests, NAEP's reporting efforts serve as an example of the kinds of materials that testing programs can make available, and the ways in which test data and information can be shared in an online setting. Stakeholders interested in NAEP data can get access to a wide range of information, and in the following sections, we provide an overview of the information included in NAEP's web

presence with respect to both content and format. These topic areas discussed are 1) the NAEP homepage, 2) interactive/media tools, 3) static, data-oriented web pages, and 4) programmatic/informational web pages. For each grouping, we offer some take-home suggestions for testing agencies tasked with developing and maintaining online assessment resources for large-scale tests. [Authors' introduction]

6. Reporting Policy and Accountability

Alexander, L., & James, H. T. (1987). *The nation's report card: Improving the assessment of student achievement* (Report of the Study Group). Washington, DC: National Academy of Education.

Recommendations of the Study Group on the National Assessment of Educational Progress (NAEP) are summarized in this report. The report is then reviewed by the National Academy of Education. The recommendations are to: (1) maintain NAEP's continuity; (2) assess the core curriculum; (3) focus on transitional grades (4, 8, and 12) and sample out-of-school 17-year-olds, adults and private school students; (4) create an independent Educational Assessment Council, with members to be appointed by the Secretary of Education; (5) provide for add-on and school district assessments; (6) assess and provide for add-on assessment of private school students; and (7) increase federal funding to 20 to 30 million dollars per year (approximately five times the present amount). The Study Group strongly recommends that achievement data be collected on each state and the District of Columbia and that state and local assessments be linked with NAEP. Curriculum areas to be tested include higher order thinking skills; reading, writing and literacy; mathematics, science, and technology; history, geography, and civics; and special topics which are assessed occasionally. Recommendations for measurement include computer assisted testing and scaling. [Authors' abstract]

DeVito, P. J. (1997). The future of the National Assessment of Educational Progress from the states' perspective. In R. Glaser, R. Linn, & G. Bohrnstedt (Eds.), *Assessment in transition: Monitoring the nation's educational progress*. Stanford, CA National Academy of Education.

Forte Fast, E. (1999, April). *Education indicators and accountability systems: Critical issues in development and reporting*. Invited panelist at the meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Forte Fast, E. (2000, January). *Rationale for selecting educational reporting indicators*. Invited presentation at the Issues in State Accountability: Making Informed Decisions about Accountability Reports workshop, Council of Chief State School Officers, San Antonio, TX.

Forte Fast, E. (2003, April). *Improving reporting and use of data in accountability systems.* Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Forte Fast, E., & Tucker, C. (2001, April). *Redesign of the student assessment reporting system in Connecticut*. Paper presented at the meeting of the American Educational Research Association, Seattle, Washington.

Glaser, R., Linn, R., & Bohrnstedt, G. (1997). *Assessment in transition: Monitoring the nation's educational progress*. Stanford, CA: National Academy of Education.

Goertz, M. E., Duffy, M. C., & Le Floch, K. C. (2001). *Assessment and accountability systems in the 50 states: 1999-2000* (Consortium for Policy Research in Education Report Series RR-046). University of Pennsylvania, Consortium for Policy Research in Education. Retrieved March 31, 2009, from http://www.cpre.org/Publications/Publications_Accountability.htm

Hamilton, L. S., & Koretz, D. M. (2002). Tests and their use in test-based accountability systems. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 13-49). Santa Monica, CA: RAND. Retrieved March 31, 2009, from https://www.rand.org/pubs/monograph_reports/MR1554/MR1554.ch2.pdf

In this chapter, we provide a short history of large-scale testing and test-based accountability, describe features of the accountability systems and tests that are in place today, and describe several ways in which state testing systems vary. We discuss content and performance standards, which in today's systems typically serve as the means for communicating a common set of goals. Following that, we present information on the features of tests, discuss several other issues related to large-scale testing, including methods of reporting, procedures for setting performance targets, and test-based rewards and sanctions. [Authors' abstract]

Jaeger, R. M. (1992). General issues in reporting of the NAEP trial state assessment results. In R. Glaser & R. L. Linn (Eds.), *Assessing student achievement in the states* (pp. 107-109). Stanford, CA: National Academy of Education.

Jaeger, R. M. (1998). *Reporting the results of the National Assessment of Educational Progress* (NAEP Validity Studies). Washington, DC: American Institutes for Research. Retrieved March 31, 2009, from http://www.air.org/publications/documents/Reporting_NAEP.pdf

This paper explores the ways results of National Assessment of Educational Progress (NAEP) data might be communicated to a variety of audiences, each with differing needs for information, interest in its findings, and sophistication in interpreting the results. The paper describes market-basket reporting as a feasible alternative to traditional NAEP reporting. Such reports would include samples of items and exercises with their scoring rubrics. The second section of the paper makes the case that in order to up-hold the strict standards of data quality, NAEP reports must format and display results to make them more accessible while discouraging readers from drawing overly broad interpretations of the data. A final section describes a detailed program of research on reporting and dissemination of NAEP findings based on these dimensions: (1) the research questions to be asked; (2) the audiences to whom the questions should be addressed; and (3) the strategies through which the questions should be pursued.

The paper suggests that the highest priority be given to research on reporting through public media, followed by making NAEP reporting more understandable and useful to school curriculum and instruction personnel, reporting to the public, and further research with state education personnel. [Author's abstract]

Jaeger, R. M. (2003). *NAEP validity studies: Reporting the results of the National Assessment of Educational Progress* (Working Paper 2003-11). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Jaeger, R. M., Gorney, B., Johnson, R., Putnam, S. E., & Williamson, G. (1994). *A consumer report on school report cards*. Kalamazoo, MI: Western Michigan University, the Evaluation Center.

Jaeger, R. M., Gorney, B., Johnson, R., Putnam, S. E., & Williamson, G. (1994). *Designing and developing effective school report cards: A research synthesis.* Kalamazoo, MI: Western Michigan University, the Evaluation Center.

Jaeger, R. M., & Tucker, C. G. (1998). *Analyzing, disaggregating, reporting, and interpreting students' achievement test results: A guide to practice for Title I and beyond*. Washington, DC: CCSSO.

Jennings, J., & Stark, D. (1997). The future of the National Assessment of Educational Progress. In R. Glaser, R. Linn, & G. Bohrnstedt (Eds.), *Assessment in transition: Monitoring the Nation's Educational Progress*. Stanford, CA: National Academy of Education.

Johnson, E., Lazer, S., & O'Sullivan, C. (1997). *NAEP reconfigured: An integrated redesign of the National Assessment of Educational Progress*. Washington, DC: National Center for Education Statistics.

Chapters in this report outline the potential plans for the redesign of the National Assessment of Educational Progress (NAEP). It is argued that any successful redesign must consider the NAEP as a whole. This report reviews overall NAEP designs and discusses the implications that each of the designs has for various functional areas. [Authors' abstract]

Koretz, D. (1995). The quality of information from NAEP: Two examples of work done in collaboration with Leigh Burstein. *Educational Evaluation and Policy Analysis, 17*(3), 280-294.

This article summarizes two research efforts, both focusing on the mathematics assessments of the National Assessment of Educational Progress, that illustrate Leigh Burstein's long- standing concern with the quality of information about the condition of education. The first examined nonresponse to NAEP test items; it found that omit rates were highest for difficult constructed-response items and that African American and Hispanic students had higher omit rates than Whites.

The second study evaluated the validity of the 1992 achievement level descriptions as characterizations of mathematics performance; it found that the descriptions and accompanying exemplar items were misleading. In response to these findings, a variety of recommendations were offered pertaining to test construction, standards setting, routine monitoring (and reporting) of data quality, and standards-based reporting of student performance. [Author's abstract]

Koretz, D. (1996). Using student assessments for educational accountability. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 171-195). Washington, DC: National Research Council.

Landgraf, K. M. (2001). *Using assessments and accountability to raise student achievement* [On-line]. Retrieved March 31, 2009, from ftp://ftp.ets.org/pub/corp/kurttest.pdf

Linn, R. L. (1998). *Assessments and accountability* (CSE Technical Report 490). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Teaching.

Uses of tests and assessments as key elements in five waves of educational reform during the past 50 years are reviewed. These waves include the role of tests in tracking and selection emphasized in the 1950s, the use of tests for program accountability in the 1960s, minimum competency testing programs of the 1970s, school and district accountability of the 1980s, and the standards-based accountability systems of the 1990s. Questions regarding the impact, validity, and generalizability of reported gains and the credibility of results in high-stakes accountability uses are discussed. Emphasis is given to three issues of currently popular accountability systems. These are (a) the role of content standards, (b) the dual goals of high performance standards and common standards for all students, and (c) the validity of accountability models. Some suggestions for dealing with the most severe limitations of accountability are provided. [Author's abstract]

Linn, R. L. (2001). Validation of the uses and interpretations of results of state assessment and accountability systems. In J. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for all students: Development, implementation, and analysis* (pp. 27-48). Mahwah, NJ: Lawrence Erlbaum Associates.

Provides an overview of validity within the context of assessment and accountability systems mandated and developed by states. Following usage in the 1999 Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council of Measurement in Education; hereafter referred to as Test Standards), the term test is used in a broad sense to include any systematic evaluative device or assessment procedure. The author uses the Test Standards as an organizing tool to discuss the types of evidence and logical arguments that those responsible for state assessment and accountability systems should develop to evaluate the

validity of the uses and interpretations that are made of results. The author begins with a brief overview of the concept of validity and the way in which thinking about the validity in the measurement system has evolved over time. This is followed by a discussion of specific uses and interpretations of results of state assessment and accountability systems and the requirements of the Test Standards to evaluate the validity of those uses and interpretations. [Author's abstract]

Linn, R. L. (in press). Accountability system design. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data in accountability testing. Yearbook of the National Society for the Study of Education,* Vol. 104, Part I.

Linn, R. L, & Dunbar, S. B. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measuremen*t, *29*(2), 177–194.

Several issues related to the design and reporting of NAEP results are discussed within the context of current expectations for NAEP and its historical origins. Procedures for establishing the content and form of assessments, including the process of developing frameworks, and eventually individual assessment items are discussed. The need to maintain a comprehensive assessment reflecting both current practice in schools and the best thinking by subject matter experts is emphasized. Issues in the design and the estimation of subpopulation parameters using conditioning variables are discussed. Finally, continuing misinterpretations of anchor item results are analyzed. [Authors' abstract]

NAEP Validity Studies Panel. (2002). *An agenda for NAEP validity research.* Washington, DC: American Institutes for Research.

National Assessment Governing Board. (1996). *Redesigning the National Assessment of Education Progress, policy statement*. Washington, DC.

National Assessment Governing Board. (2006). *Policy statement on reporting, release, and dissemination of NAEP results.* Retrieved March 31, 2009, from http://www.nagb.org/policies/PoliciesPDFs/Reporting%20and%20Dissemination/Reporting,%20Release,%20and%20Dissemination%20of%20NAEP%20Results.pdf

National Center for Education Statistics. (2000, May). *Rewards for NAEP: Proposals and consequences*. Paper prepared for meeting of the Design and Analysis Committee , Washington, DC.

National Education Goals Panel. (1991). *The national education goals report: Building a nation of leaders.* Washington, DC.

Snodgrass, D., & Salzman, J. A. (2002, April). *Creating the Rosetta stone: Deciphering the language of accountability to improve student performance.* Paper presented

at the meeting of the American Educational Research Association, New Orleans, LA.

The objectives of this paper are threefold. First, a model is proposed for unifying massive amounts of conceptual and numerical information flowing from the measures of the accountability movement in Ohio and the materials that are publicly available to educators. Second, this model is translated into useable forms of information that help teachers drive instructional practices in their classrooms. Third, the efficacy of this model meant to improve state-mandated proficiency scores at the district level is discussed. The paper attempts to integrate numerical, pictorial, graphical and narrative information about the Ohio Proficiency Tests in a way that provides the reader with a rudimentary model of an educational Rosetta Stone. This tool helps educators decipher the contents of the Ohio proficiency tests at a level complex enough so educators can identify basic and fundamental instructional needs of their students. Six appendixes contain sample items from the Ohio proficiency tests. [Authors' abstract]

Twing, J. S. (2008, March). *Score reporting, off-the-shelf assessments and NCLB: Truly an unholy trinity*. Paper presented at the meeting of National Council on Measurement in Education, New York, NY.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2007). *Comprehensive evaluation of NAEP: Utility study final report* (Final Report to Congress). [Also Center for Educational Assessment Report No. 624]. Amherst, MA: University of Massachusetts, Center for Educational Assessment.

7. Sample reports

Individual reports

British Columbia Ministry of Education. (2009). *Foundation skills assessment: Individual student results report.* Victoria, BC: Author. Retrieved March 31, 2009, from http://www.bced.gov.bc.ca/assessment/fsa/results_interpret.htm

The FSA Individual Student Results Report sent home from schools are provided in English. To help families and students for whom English is not the language spoken in the home, translated versions of the Individual Student Results Report are available on this site in 14 additional languages. [Author's abstract]

Connecticut State Board of Education. (2009). *Individual student performance reports*: *Connecticut mastery test, 4th generation.* Hartford, CT: Author. Retrieved March 21, 2009, from http://www.ctreports.com/

CTB/McGraw-Hill. (2003). *Interactive reports page* [On-line]. Retrieved March 31, 2009, from http://www.ctb.com/mktg/terranova/tn_reports_noflash.jsp

Education Quality and Accountability Office. (2006). *Individual student report: Grade 6 assessment of reading, writing and mathematics, 2005-2006.* Toronto, ON: Author. Retrieved March 31, 2009, from http://www.gecdsb.on.ca/parents/eqao/Gr6SampleISR.pdf

Education Quality and Accountability Office. (2007). *Individual student report: Assessment of reading, writing and mathematics, primary division (grades 1-3), 2006-2007.* Toronto, ON: Author. Retrieved March 31, 2009, from http://www.eqao.com/pdf_e/07/07P082e.pdf

Harcourt Educational Measurement. (2002). *Stanford achievement test series, tenth edition: Sample reports*. San Antonio, TX: Author. Retrieved March 31, 2009, from http://pearsonassess.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=SAT10C&Mode=scoring&Leaf=SAT10C_1

Illinois State Board of Education. (2008). *Spring 2008 PSAE individual student report*. Springfield, IL: Author. Retrieved March 31, 2009, from http://www.isbe.state.il.us/assessment/pdfs/PSAE_ISR_Sample.pdf.

Minnesota Department of Education. (2008). *Minnesota comprehensive assessments – series II: Student report.* Minneapolis, MN: Author. Retrieved March 31, 2009, from http://www.cloquetcommunityed.com/files/filesystem/GRAD%20Sample%20Report.pdf

Mitchell, K. J., & Haynes, R. (1990). Score reporting for the 1991 medical college admission test. *Academic Medicine, 65*(12), 719-23.

Data used in a major review of the system for reporting scores on the Medical College Admission Test (MCAT) are presented and discussed. The data demonstrated the value of the current score-reporting system and led to retention of the 15-point MCAT score scale in 1991. [Authors' abstract]

University of Iowa. (2001). *The Iowa tests: Report to students and parents.* Itasca, IL: Riverside Publishing.

Group reports

Campbell, J. R., Voekl, K. E., & Donahue, P. L. (1997). *NAEP 1996 trends in academic progress: Achievement of U.S. students in science, 1969 to 1996; mathematics, 1973 to 1996; reading, 1971 to 1996; and writing, 1984 to 1996* (NCES Report No. 97-985). Washington, DC: U.S. Department of Education.

Connecticut State Board of Education. (2009). *Public summary performance reports: Connecticut Mastery Test, 4th generation.* Hartford, CT: Author. Retrieved March 21, 2009, from http://www.ctreports.com/

Donahue, P. L., Voekl, K. E., Campbell, J. R., & Mazzeo, J. (1999). *The NAEP 1998 reading report card for the nation* (NCES 1999-459). Washington DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Educational Testing Service. (1997). *NAEP 1996 mathematics report for Milwaukee public schools grade 8: Findings from a special study of the National Assessment of Educational Progress*. Princeton, NJ: Author.

Gonzales P., Guzmán, J. C., Partelow, L., Pahlke, E., Jocelyn, L., Kastberg, D., & Williams, T. (2004). *Highlights from the Trends in International Mathematics and Science Study (TIMSS) 2003* (NCES 2005–005). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., & Brenwald, S. (2008). *Highlights from TIMSS 2007: Mathematics and science achievement of U.S. fourth- and eighth-grade students in an international context* (NCES 2009–001). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Grigg, W., Lauko, M., & Brockway, D. (2006). *The nation's report card: Science 2005* (NCES 2006-466). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Iowa Testing Programs. (2009). *2008-2009 score reports and norms.* Iowa City, IA: University of Iowa, College of Education. Retrieved March 31, 2009, from http://www.education.uiowa.edu/itp/documents/08-09ScoreReportsNorms.pdf

Lee, J., Grigg, W., & Dion, G. (2007). *The nation's report card: Mathematics 2007* (NCES 2007–494). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Lee, J., Grigg, W., & Donahue, P. (2007). *The nation's report card: Reading 2007* (NCES 2007–496). National Center for Education Statistics, Institute of

Education Sciences, U.S. Department of Education, Washington, DC.

Lee, J., & Weiss, A. (2007). *The nation's report card: U.S. History 2006* (NCES 2007–474). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Louisiana Department of Education. (2008). *2008 LEAP/GEE subgroup reports*. Baton Rouge, LA: Author. Retrieved March 31, 2009, from http://www.doe.state.la.us/lde/saa/1339.html

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., O'connor, K. M., Chrostowski, S. J., & Smith, T. A. (2000). *TIMSS 1999 international mathematics report: Findings from IEA's repeat of the Third International Mathematics and Science Study at the eighth grade.* Chestnut Hill, MA: Boston College.

National Center for Education Statistics. (2001a). *The nation's report card: Mathematics highlights 2000.* Washington, DC: U.S. Department of Education.

National Center for Education Statistics. (2001b). *The nation's report card: Fourth-grade reading highlights 2000.* Washington, DC: U.S. Department of Education.

Northern Illinois University. (2009). *Interactive Illinois report card*. Dekalb: IL. Retrieved March 31, 2009, from http://iirc.niu.edu/Default.aspx

Perie, M., Grigg, W., & Dion, G. (2005). *The nation's report card: Mathematics 2005* (NCES 2006–453). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Shaughnessy, C. A., Nelson, J. E., & Norris, N. A. (1997). *NAEP 1996 Mathematics cross-state data compendium for the grade 4 and grade 8 assessment*. Washington, DC: National Center for Education Statistics.

Weiss, A. R., Lutkus, A. D., Hildebrant , B. S., & Johnson, M. S. (2001). *The nation's report card: Geography 2001* (NCES 2002–484). U.S. Department of Education, National Center for Education Statistics. Washington, DC: Office of Educational Research and Improvement.

Interpretive guides

College Board, & National Merit Scholarship Corporation (2008). *Understanding 2008 PSAT/NMSQT scores*. New York: College Board. Retrieved March 4, 2009, from http://professionals.collegeboard.com/profdownload/understanding-psat-nmsqt-scores.pdf

CTB/McGraw-Hill. (2003). *Beyond the numbers: A guide to interpreting and using the results of standardized achievement tests*. Monterey, CA: Author. Retrieved March 31, 2009, from http://www.ctb.com/articles/article_information.jsp?CONTENT%3C%3Ecnt_id=10134198673323311&FOLDER%3C%3Efolder_id=2534374302134983&bmUID=1235418182990

A sound assessment and evaluation program reinforces educational decision and can positively impact individual students as well as parents or guardians, classes, schools and communities. [Author's abstract]

Delaware Department of Education. (2008). *Delaware student testing program: A score results guide for parents.* Dover, DE: Author. Retrieved March 31, 2009, from http://www.doe.k12.de.us/aab/files/2008%20Parents%20Guide.pdf

Education Quality and Accountability Office. (2008). *Using data to promote student success: A brief guide to assist school administrators in interpreting their data.* Toronto, ON: Author. Retrieved March 31, 2009, from http://www.eqao.com/pdf_e/08/Dudg_xe_0608_web.pdf

Horkay, N. (1999). *The NAEP Guide* (NCES 2000—456). Washington, DC: National Center for Education Statistics.

Louisiana Department of Education. (2008). *Interpretive guide: Grades 4, 8, 10, and 11 criterion-referenced tests*. Baton Rouge, LA: Author. Retrieved March 31, 2009, from http://www.doe.state.la.us/lde/uploads/1278.pdf

Massachusetts Department of Education. (2008). *The Massachusetts comprehensive assessment system: Guide to the 2008 MCAS for Parents/Guardians.* Boston, MA: Author. Retrieved March 31, 2009, from http://www.doe.mass.edu/mcas/2008/results//english.pdf

Massachusetts Department of Education. (2008). *The Massachusetts comprehensive assessment system: Guide to interpreting the Spring 2008 MCAS reports for schools and districts.* Boston, MA: Author. Retrieved March 31, 2009, from http://www.doe.mass.edu/mcas/2008/results/interpretive_guide.pdf

Minnesota Department of Education. (2008). *Minnesota Interpretive Guide (2007-2008)*. Roseville, MN: Author. Retrieved March 31, 2009, from http://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/Assessments/MCA/Reports/index.html.

New Jersey Department of Education. (2005). *October 2005 and March 2006 high school proficiency assessment (HSPA): Cycle I and cycle II score interpretation manual.* Trenton, NJ: Author. Retrieved March 31, 2009, from http://www.state.nj.us/education/assessment/hs/hspa/info/

North Carolina Department of Public Instrument. (2007). *Understanding Student Achievement within the ND State Assessment: A Primer.* Retrieved March 31, 2009, from http://www.dpi.state.nd.us/testing/assess/understand0406.pdf

Iowa Testing Programs. (2003). *Interpretive guide for the achievement levels report. ITBS/ITED testing program.* Iowa City, IA: University of Iowa, College of Education. Retrieved March 31, 2009, from http://www.education.uiowa.edu/itp/downloads.aspx

Virginia Department of Education. (2002). *Virginia standards of learning assessments: Understanding your child's SOL report.* Richmond, VA: Author.

State test report websites

Knupp, T., & Ansley, T. (2008, March). *Online, state-specific assessment score reports and interpretive guides.* Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.

Alabama	
Department:	http://www.alsde.edu/html/home.asp
Score Reports:	http://www.alsde.edu/html/reports_menu.asp
Interpretive Guide:	http://www.alsde.edu/html/sections/doc_download.asp?section=100&id=5310&sort
Alaska	
Department:	http://www.eed.state.ak.us/
Score Reports:	http://www.eed.state.ak.us/tls/assessment/results/results2007.html
Interpretive Guide:	http://www.eed.state.ak.us/tls/assessment/sba/Spring07/GTIs/AK-Gr8-PS-GTI_LTR.pdf
Arizona	
Department:	http://www.ade.az.gov/
Score Reports:	http://www.ade.az.gov/profile/publicview/
Interpretive Guide:	http://www.ade.az.gov/standards/downloads/AIMSDPAcolor.pdf
Arkansas	
Department:	http://arkansased.org/
Score Reports:	http://arkansased.org/testing/test_scores.html
Interpretive Guide:	http://arkansased.org/testing/pdf/rig_benchmark_spr07.pdf
California	
Department:	http://www.cde.ca.gov/
Score Reports:	http://star.cde.ca.gov/star2007/viewreport.asp
Interpretive Guide:	http://www.cde.ca.gov/ta/tg/sr/documents/guides07tests.pdf
Colorado	
Department:	http://www.cde.state.co.us/
Score Reports:	http://www.cde.state.co.us/cdeassess/documents/csap/usa_index.html
Interpretive Guide:	http://www.cde.state.co.us/cdeassess/documents/csap/2006/GR3-8GTI_CSAP2006.pdf

Connecticut	
Department:	http://www.sde.ct.gov/sde/site/default.asp
Score Reports:	http://www.cmtreports.com/
Interpretive Guide:	http://www.csde.state.ct.us/public/cedar/assessment/cmt/resources/misc_cmt/2007_cmt_interpretive_guide.pdf
Delaware	
Department:	http://www.doe.state.de.us/
Score Reports:	http://dstp.doe.k12.de.us/DSTPmart9/
Interpretive Guide:	http://www.doe.k12.de.us/aab/2007%20Parents%20Guide.pdf
Florida	
Department:	http://www.fldoe.org/
Score Reports:	http://fcat.fldoe.org/fcatscor.asp
Interpretive Guide:	http://fcat.fldoe.org/pdf/ufr_07_content.pdf
Georgia	
Department:	http://www.k12.ga.us/
Score Reports:	http://www.k12.ga.us/DMGetDocument.aspx/2007%20CRCT%20Testing%20Brief.pdf?p=6CC6799F8C1371F69366A7A584FF130D7FDA40BCAD02B18654BF283BF13F8753&Type=D
Interpretive Guide:	NONE
Hawaii	
Department:	http://doe.k12.hi.us/
Score Reports:	http://arch.k12.hi.us/PDFs/nclb/2006/HSA%202006%20STATE%20RESULTS%20PresentationRAM.pdf
Interpretive Guide:	http://arch.k12.hi.us/PDFs/nclb/2008/Guide%20to%20the%202008%20HSA%20and%20AYP,%20rc%2012-03-07.pdf
Idaho	
Department:	http://www.sde.idaho.gov/
Score Reports:	http://www.boardofed.idaho.gov/saa/ISAT_FA07.asp
Interpretive Guide:	http://www.boardofed.idaho.gov/saa/documents/IdahoParentBrochure-OSBE.pdf
Illinois	
Department:	http://www.isbe.state.il.us/
Score Reports:	http://www.isbe.state.il.us/research/htmls/test_results.htm#isat
Interpretive Guide:	http://www.isbe.state.il.us/assessment/pdfs/ISAT_Interpr_Guide_2007.pdf
Indiana	
Department:	http://www.doe.state.in.us/
Score Reports:	http://www.doe.state.in.us/istep/2007/welcome.html
Interpretive Guide:	http://www.doe.state.in.us/istep/pdf/GTI/49936-W_GTI_F07IN.pdf
Kansas	
Department:	http://www.ksde.org/
Score Reports:	http://online.ksde.org/rcard/summary/state.pdf
Interpretive Guide:	http://conferences.ksde.org/LinkClick.aspx?fileticket=sbNt7ihv7%2fU%3d&tabid=1334&mid=2377
Kansas	
Department:	http://www.education.ky.gov/KDE/
Score Reports:	http://www.education.ky.gov/NR/rdonlyres/942E5B6D-2227-4DB3-

	881B-12FA0FA55418/0/MediaContentReport2007.pdf
Interpretive Guide:	http://www.education.ky.gov/NR/rdonlyres/502361D9-6DFE-41A4-A8F4-94084B86B7A7/0/2005CATSInterpretiveGuideV31.doc
Louisiana	
Department:	http://www.doe.state.la.us/lde/index.html
Score Reports:	http://www.doe.state.la.us/lde/saa/1337.html
Interpretive Guide:	http://www.doe.state.la.us/mark/lde/uploads/1278.pdf
Maine	
Department:	http://www.maine.gov/education/
Score Reports:	http://www.maine.gov/education/mea/0607meascores/0607_0506_state_results.html
Interpretive Guide:	http://mainegov-images.informe.org/education/mea/techmanual0506.pdf
Maryland	
Department:	http://marylandpublicschools.org/MSDE
Score Reports:	http://www.mdreportcard.org/Assessments.aspx?WDATA=State&K=99AAAA
Interpretive Guide:	http://www.mdk12.org/data/explorer/index_b.html
Massachusetts	
Department:	http://www.doe.mass.edu/
Score Reports:	http://www.doe.mass.edu/mcas/results.html
Interpretive Guide:	http://www.doe.mass.edu/mcas/2007/pgguide/english.pdf
Michigan	
Department:	http://www.michigan.gov/mde
Score Reports:	http://www.michigan.gov/mde/0,1607,7-140-22709_31168_40135---,00.html
Interpretive Guide:	http://www.michigan.gov/documents/mde/F07_Guide_to_Reports_sm_223910_7.pdf
Minnesota	
Department:	http://education.state.mn.us/mde/index.html
Score Reports:	http://education.state.mn.us/MDE/Data/Data_Downloads/Accountability_Data/Assessment_MCA_II/MCA_II_Excel_files/index.html
Interpretive Guide:	http://education.state.mn.us/MDE/groups/assessment/documents/publication/031107.pdf
Mississippi	
Department:	http://www.mde.k12.ms.us/
Score Reports:	http://orsap.mde.k12.ms.us:8080/MAARS/indexProcessor.jsp
Interpretive Guide:	http://www.mde.k12.ms.us/acad/osa/gltp.html
Missouri	
Department:	http://dese.mo.gov/
Score Reports:	http://dese.mo.gov/divimprove/assess/Missouri%20Assessment%20Program%20State%20Board%202007.ppt
Interpretive Guide:	http://dese.mo.gov/divimprove/assess/2007_gir_manual.pdf
Montana	
Department:	http://www.opi.mt.gov/
Score Reports:	http://data.opi.state.mt.us/irisreports/
Interpretive Guide:	http://www.opi.mt.gov/Assessment/Phase2.html

Nevada	
Department:	http://www.nde.state.nv.us/
Score Reports:	http://www.nevadatestreports.com/NevadaCode/SelectionsMenu.aspx
Interpretive Guide:	NONE
New Hampshire	
Department:	http://www.ed.state.nh.us/education/
Score Reports:	http://reporting.measuredprogress.org/nhprofile/reports.aspx?view=11
Interpretive Guide:	http://reporting.measuredprogress.org/nhprofile/documents/necap/Guide%20to%20Using%20the%202007%20NECAP%20Reports.pdf
New Jersey	
Department:	http://www.state.nj.us/education/
Score Reports:	http://www.state.nj.us/education/assessment/ms/
Interpretive Guide:	http://www.state.nj.us/education/assessment/ms/gepa_guide.pdf
New Mexico	
Department:	http://www.ped.state.nm.us/
Score Reports:	http://www.ped.state.nm.us/AssessmentAccountability/AcademicGrowth/NMSBA.htm
Interpretive Guide:	NONE
New York	
Department:	http://www.nysed.gov/
Score Reports:	http://www.emsc.nysed.gov/irts/
Interpretive Guide:	http://www.nysparents.com/pdfs/nys_NYSTP_2007_M_english.pdf
North Carolina	
Department:	http://www.dpi.state.nc.us/
Score Reports:	http://www.dpi.state.nc.us/docs/accountability/testing/reports/green/0506Greenbook.pdf
Interpretive Guide:	http://www.dpi.state.nc.us/docs/accountability/grade_8parenteacherreport_final.pdf
North Dakota	
Department:	http://www.dpi.state.nd.us/
Score Reports:	http://www.dpi.state.nd.us/testing/assess/data/achieve0607m.pdf
Interpretive Guide:	http://www.dpi.state.nd.us/testing/assess/understand0406.pdf
Ohio	
Department:	http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDefaultPage.aspx?page=1
Score Reports:	http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=263&ContentID=15606&Content=40999
Interpretive Guide:	http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=222&ContentID=17597&Content=36891
Oklahoma	
Department:	http://www.sde.state.ok.us/home/defaultie.html
Score Reports:	http://www.sde.state.ok.us/studentassessment/pdfs/testresults07.pdf
Interpretive Guide:	http://www.sde.state.ok.us/studentassessment/06-07/Grades%203-8%20TIM%202007.pdf
Oregon	
Department:	http://www.ode.state.or.us/

Score Reports:	http://www.ode.state.or.us/data/schoolanddistrict/testresults/reporting/PublicRpt.aspx
Interpretive Guide:	http://www.ode.state.or.us/teachlearn/testing/manuals/2007/asmttechmanualvol6_interpguide.pdf
Pennsylvania	
Department:	http://www.pde.state.pa.us/
Score Reports:	http://www.pde.state.pa.us/a_and_t/lib/a_and_t/2007_State_Level_PSSA_Results.pdf
Interpretive Guide:	http://www.pde.state.pa.us/a_and_t/lib/a_and_t/2003MathandReadingHBforReportInterpretation.pdf
Rhode Island	
Department:	http://www.ride.ri.gov/
Score Reports:	http://www.ride.ri.gov/Assessment/Results.aspx
Interpretive Guide:	http://www.ride.ri.gov/Assessment/DOCS/NECAP/2006_ReportsInterp_Guide.pdf
South Carolina	
Department:	http://ed.sc.gov/
Score Reports:	http://ed.sc.gov/topics/assessment/scores/pact/2007/statescoresdemo.cfm
Interpretive Guide:	http://ed.sc.gov/agency/offices/assessment/pact/documents/PACTUserGuide07BlackWhite.pdf
South Dakota	
Department:	http://doe.sd.gov/
Score Reports:	https://sis.ddncampus.net:8081/nclb/index.html
Interpretive Guide:	http://doe.sd.gov/octa/assessment/docs/DakotaSTEPInterpretiveGuide.pdf
Tennessee	
Department:	http://www.tennessee.gov/education/
Score Reports:	http://www.k-12.state.tn.us/rptcrd05/
Interpretive Guide:	http://www.state.tn.us/education/assessment/doc/Form_R_Parent.pdf
Texas	
Department:	http://www.tea.state.tx.us/sboe/
Score Reports:	http://www.tea.state.tx.us/student.assessment/reporting/
Interpretive Guide:	http://www.tea.state.tx.us/student.assessment/resources/guides/parent_csr/2008/TK08_Apr_ParentBroch_G8_M.pdf
Utah	
Department:	http://www.schools.utah.gov/
Score Reports:	http://www.schools.utah.gov/assessment/documents/Results_CRT_State_05-07.pdf
Interpretive Guide:	NONE
Vermont	
Department:	http://education.vermont.gov/
Score Reports:	http://education.vermont.gov/new/html/pgm_assessment/data.html#necap
Interpretive Guide:	http://education.vermont.gov/new/pdfdoc/pgm_assessment/necap/reporting_workshops_07/using_reports.pdf

Virginia	
Department:	http://www.doe.virginia.gov/
Score Reports:	https://p1pe.doe.virginia.gov/reportcard/report.do?division=All&schoolName=All
Interpretive Guide:	NONE
Washington	
Department:	http://www.sbe.wa.gov/
Score Reports:	http://reportcard.ospi.k12.wa.us/WASLCurrent.aspx?schoolId=1&reportLevel=State&year=2006-07&orgLinkId=&waslCategory=1&gradeLevelId=8&chartType=1#&gradeLevel=8
Interpretive Guide:	NONE
West Virginia	
Department:	http://wvde.state.wv.us/
Score Reports:	http://westest.k12.wv.us/2007reports.html
Interpretive Guide:	http://westest.k12.wv.us/pdf/westestguidetointerpertation.pdf
Wisconsin	
Department:	http://dpi.state.wi.us/
Score Reports:	http://www2.dpi.state.wi.us/wsas/statewkce.asp
Interpretive Guide:	http://dpi.state.wi.us/oea/pdf/adminguide07.pdf
Wyoming	
Department:	http://www.k12.wy.us/
Score Reports:	http://www.k12.wy.us/SAA/Paws/PAWS07/state_07.asp
Interpretive Guide:	NONE