Demands on Users for Interpretation of Achievement Test Scores: Implications for the Evaluation Profession

Gabriel Mario Della-Piana
Independent Consultant

Michael Gardner
University of Utah

Background: Professional standards for validity of achievement tests have long reflected a consensus that validity is the degree to which evidence and theory support interpretations of test scores entailed by the intended uses of tests. Yet there are convincing lines of evidence that the standards are not adequately followed in practice, that standards alone are not sufficient guides to action, and that reviewers of tests do not call attention to important kinds of validity evidence that might support the demanding process of making sense of test scores or reasoning from test scores.

Purpose: The intent of this article is to make more transparent the demands of achievement test interpretation on users in instructional contexts and to open up a dialogue on implications for the evaluation profession for improvement of practice along lines already set out by evaluation theorists.
Setting: Not applicable.

Intervention: Not applicable.

Research Design: Not applicable.

Data Collection and Analysis: Review of current practice.

Findings: The article makes transparent the lack of attention to validation of achievement tests to support inferences relevant to intended uses in instruction and project evaluation. Elements of a model for the process of reasoning from test scores are articulated. The cognitive demands on the test score user are illustrated in achievement test contexts in writing, science, and mathematics. Implications are drawn for deliberation on issues and for the development of casebooks to guide practice.

Keywords: *assessment; test validation; test users; test interpretation*

The intent of this paper is to make transparent the demands of achievement test interpretation on users and to open up a dialogue on implications for the evaluation profession. The focus on demands for interpreting achievement tests is based in part on the widespread use of tests, not only for both instructional and accountability decisions, but for the interpretation of findings in evaluation of interventions. It is the study of interpretations of findings in the current context that is important for improvement of practice along lines already set out by theorists.

It would be reasonable to expect that appropriate inference from and use of student achievement test scores for instructional decisions and program evaluation practice might be strongly facilitated, if not assured, by test developers and users being guided by consensual published professional standards in testing (Standards for Educational and Psychological Testing published by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education) (AERA, APA, NCME, 1999, under revision, hereafter, Standards). There are signs however that the Standards are not adequately followed in practice or are not sufficient to guide practice.

Some signs of lack of compliance with the Standards come from responses of the measurement community following the 2005 call for comments on the revision of 1999 Standards. Wise (2006) expresses concerns about the lack of compliance due to the complexity of who is the developer (now often a collaboration between test publishers and purchasers and users), the lack of sufficient number of trained professionals in measurement (Herszenhorn, 2006), and the lack of enforcement mechanisms, or even reviews, of current tests. Wise proposes incentives coupled with providing the assistance and expertise needed. Linn (2006) argues that instead of a revision of current Standards, the considerable effort and resources might be better spent in producing companion standards dealing with special cases of application and casebooks that illustrate how the Standards apply to specific contexts of use. Nichols and Williams (2009) summarize views on who is responsible for collecting evidence and suggest conditions under which that responsibility rests with the developer(s) or the users of tests. Nichols and Wiliams cite Linn (1998) on the complexity of specifying the user, given the accountability and funding mechanisms under which the federal government is a user (requiring certain practices for funding), the school district is a user (requiring performance under accountability regulations for both practice and results), and the state legislature is a user through legislation. The issue of responsibility is also taken up by Madaus, Russell, and Higgins (2009) under the topic of why and how high stakes tests should be monitored (p. 197).

The interpretive burden and importance of making sense of test scores is increased with federal funding for development of new testing systems and new kinds of tests (U. S. Department of Education, 2011, March). These initiatives, encompassing multiple measures and open-ended assessment, open up a greater need for understanding the demands of integrating several lines of evidence into an interpretation and decision. While this is the case for individual student achievement, the integration of multiple lines of evidence into an interpretation is a strong part of Scriven's evaluation model in the *Key Evaluation Checklist* (KEC) (Scriven, 2001, 2007; Davidson, 2005) and philosophical treatment of practice in interpretation of multiple lines of evidence (Schwandt, 1998, 2008a, 2008b).

The paper proceeds as follows. Terminology: Testing, Assessment, and Validity; The Process of Reasoning from Test Scores; Illustrative Demands On The User For Reasoning From Multiple Sources Of Evidence; Contributions to Reasoning from Test Scores Made by Reviewers of Educational Tests and Assessments; Implications for a Line of

Practical Action by the Evaluation Profession.

Terminology: Testing, Assessment, and Validity

In order to argue for more professional attention to issues around validity of inferences made from test scores, one needs some agreement on key technical terms. This is especially so since testing terms are used in different ways in the literature. For example, *The Student Evaluation Standards: How to Improve Evaluations of Students* (Joint Committee on Standards for Evaluation, 2001) define assessment as the process of collecting information about a student. *The Program Evaluation Standards* (Joint Committee on Standards for Evaluation, 2011) define assessment as determining "the relative or absolute position on some variable of interest." Thus, for clarity of exposition, the authors follow definitions in the Standards (AERA, APA, NCME, 1999) as briefly noted here.

A test is "an evaluative device or procedure in which a sample of...behavior is obtained and...evaluated and scored using a standardized process" (p.3). Assessment is "a process that integrates test information with information from other sources" including other tests, individual history, and context (p. 3). "Validity is...the most fundamental consideration in developing and evaluating tests" (p. 9). Though there appears to be a fragile consensus on some parts of the Standards (Lissitz, 2009), it is generally agreed that validity, as defined there, is "...the degree to which evidence and theory support interpretations of test scores entailed by proposed uses of tests" (Standards, p.9). This definition immediately implicates all participants in the testing process as responsible parties for influencing validity since valid interpretations clearly depend on design, development, purchasing, training, and use of tests. "A sound validity argument integrates various strands of evidence and theory to support the intended interpretation of test scores for specific uses" (p. 17). Those strands of evidence, drawn upon as appropriate, include: evidence of representativeness of test content; the response processes of the test taker; test-criterion evidence; convergent and discriminant validity evidence (i.e., patterns of association between and among scores consistent with theoretical expectations); validity generalization (the degree to which evidence of test-criterion relations can be generalized to a new situation without further study); and the consequences of testing on the student, the teacher and the educational system. These are all sources of evidence to be considered for evaluating a proposed interpretation of a test for a particular purpose whether one considers validity as a unitary construct (the current consensus) or considers the sources of evidence as separate. The Standards cautions that, "If a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary" (Standard 1.4). The Standards also cautions with respect to "construct underrepresentation" referring to a test failing "to capture important aspects of the construct" that the user intended to capture with the test. These two cautions are especially relevant to the evaluator who often for reasons of budget or otherwise uses the extant achievement test for evaluating a program where the program construct includes characteristics not tapped by the extant test (p. 10). In other words this kind of lack of validity evidence challenges the evaluator with respect to gathering other evidence and taking the construct underrepresentation into account in interpretations of findings.

Since we will take the view that validity resides in the interpretation of the test, it is important to ask what that interpretation (the process of reasoning from test scores) might look like.

The Process of Reasoning from Test Scores

The process of reasoning from a variety of sources of evidence in making an interpretation of test scores along with other information about a test taker and test context is portrayed by Pellegrino, Chudowsky, and Glaser (2001) in a National Research Council report as an assessment triangle in which the corners of the triangle represent the key elements underlying any assessment. Those elements are: (a) a model of student cognition and learning in a given domain, (b) beliefs about the kinds of observation that are required to provide evidence of student competencies, and (c) an interpretation process for making sense of the evidence. Lohman and Nichols (2006), commenting on the NRC report, note that for even the best testing practices to have the desired impact, further work needs to be done in building bridges and training for the next generation of cognitive scientists and psychometricians (and we would add teachers, evaluators, and trainers of these professionals). Also, beyond the triangle, assessment must address the student as nested in ecological layers within herself, the classroom, the school, the family, the community, and the current political, social, and economic system (Berliner, 2005; Bronfrenbrenner, 1979).

The assessment triangle is a particularly instructive guide for the educational evaluator who focuses on program outcomes as the target of “what is to be observed” and “what dimensions of student cognition are relevant” beyond test scores alone. While not using the terms in the assessment triangle, the “strands of evidence and theory” to be considered by an evaluator as targets of observation are illustrated in Brickell’s 1976 paper recently reprinted in the *Journal of MultiDisciplinary Evaluation* (Brickell, 1976, 2011). A few phrases from the article tell much about what to measure in addition to program intents and test scores: “target hit, not target set” (don’t just assess what was intended or you might miss what is being accomplished); “instructional processes rather than stated goals” (how are performers really doing what you are asking them to do); the variables which the [program] theory indicates as crucial to the program being actually operationalized (what does theory tell you should be happening that current goals or objectives do not); and measuring what you can see [when you observe a program in action or interview or survey participants].

Without Brickell’s hints for finding what to measure, and the assessment triangle, interpretations will be faulty and impoverished. For example, in a fifth grade intervention to improve writing performance, students were paired randomly to provide feedback to each other on early drafts. Informal classroom observation revealed the following. A low performing male student (let’s call him Zachary) and higher performing female student (let’s call her Melinda) were paired. At one point in commenting, Zachary pointed to Melinda's paper and said, "You say this here and you say this here". Melinda responded, "Yes, that's why I started a new paragraph". Zachary repeated, "But you say this here and this here." After several more similar interactions the teacher asked Melinda to read what was pointed to. Melinda read it, her eyes opening wide, and smiling responded, "transition, transition, you caught my missing transition." The writing test scores might not change in the short term for Zachary, but the cognitive discriminations that were made visible, Melinda's realization that others not on par with her speed of learning may be able to inform her work, and Zachary's metacognitive awareness that he "knows something" all contribute to a rich interpretation of writing test scores in this context.

Finally, the evaluation profession has not been silent on guides to a process of reasoning from test scores in the case of multiple measures. The integration of multiple sources of evidence in evaluation has been articulated in Scriven's KEC (1991, 2003) checkpoint eleven on significance (synthesis). It has been elaborated by Davidson (2005) taking off from Scriven's KEC checkpoint eleven relabeled "overall significance." The demands on the user of evaluation information (tests and other information) for sound interpretation become clear in those sources.

A brief paraphrasing of the characterization of checkpoint 11 (overall significance of conclusions) as described by Davidson (2005) further clarifies the demands of interpretation on the person drawing conclusions from data. It is an evaluation team that does the interpretation, or in this case summary and synthesis. The synthesis draws on checkpoint 6 (process evaluation—value and efficiency of content and implementation), checkpoint 7 (outcome evaluation--impact and value of impacts), checkpoint 8 & 9 (comparative cost-effectiveness—cost to all impacted compared with alternative uses of available resources including acceptability and value to all), and checkpoint 10 (exportability—is the concept, design, or approach of the evaluand exportable or have value or utility in other places). The point is that just as with an individual student test score, the evaluator's interpretation of all findings with respect to an evaluand is a demanding task and responsibility that should itself be studied and especially noted that valuing or values run through the entire process.

The collection of a range of information and synthesis or reasoning from that information is recognized as so demanding and subject to the evaluator's biases that a separate evaluation of the evaluation (focusing on synthesis) has been promoted to bring out important issues with respect to interpretation and misinterpretation not considered by the evaluator. Scriven has taken the KEC in that direction labeling it "meta-evaluation" (Scriven, 1991, 2009), as has Davidson (2005). See also, Stufflebeam, (2011) on meta-evaluation.

Further Illustrative Demands on the User for Reasoning from Multiple Sources of Evidence

Cognitive and other demands on the test user for assessment (integrating multiple sources of relevant information into an interpretation) are illustrated with three concrete examples: a new type of science item, writing assessment, and a number series item.

The Demands on Interpretation of Scores on New Types of Science Items

In a discussion of science assessment under the National Assessment of Educational Progress (NAEP), Fu, Raizen, and Shavelson (December, 2009) comment on the new framework for assessment going beyond old item types to allow students to represent the structure of their knowledge. For example, students are asked “to draw and describe connections among science concepts “, producing what are called “concept maps”. Thus, test-takers use “labeled, directional arrows to link pairs of concepts (e.g., chloroplasts, green plants, and photosynthesis) and explain the relations among them (e.g., ‘requires’ or ‘contains’)”, p. 1637. Fu and others contend that for this kind of test score, unless performance is reported in enough detail, and adequately, it is difficult to make valid interpretations. That need for adequate reporting on new types of items such as concept maps is further supported in a volume on, Knowing what students know (Pellegrino, Chudowsky, & Glaser, 2001). For example, students were interviewed to explain their thinking behind their responses on “concept maps” in order to reveal the structure of their knowledge and it was found that the scoring of concept maps overstated student understanding and reasoning (pp. 209-211). With new types of items the challenge of interpretation is increased as this example demonstrates.

The Demands on the User for Interpreting Writing Assessment Scores

An example from writing assessment provides another glimpse into the complexity of interpreting test scores. Della-Piana (2008) reports on a case study on interpreting a writing test score. One simple example from the study illustrates the demands on the test user. Two students get similar scores on writing conventions with different patterns of errors that are not identified in the score report, placing a diagnostic burden on the user. Both students get low scores (1 or 2 on a 4-point scale) for problems with “conventions” (spelling, punctuation, capitalization, paragraphing). One student, when asked to read her unpunctuated, misspelled writing, reads it with proper intonation, fluency, and correct punctuation of misspelled words (such as aprtment, hgren poroxside, droped, scrache, chaut for caught, nekt for next). The other student does not perform well on those skills. The student who reads even her own misspelled words correctly and reads with fluency and proper intonation can pick up skills on “conventions” and spelling easily with a little guidance, while the other student would need more direct instruction with an instructor or aide. Other test information on the fluent reading student included scores of, beyond the 90th percentile, on reading comprehension and language expression; 86th percentile on listening comprehension; and 50th percentile on language mechanics. Integrating all this information with knowledge of language development adds to the diagnostic demands on the teacher and the evaluator and suggests that the evaluator must have someone on the team that has expertise in language development and assessment to aid in decisions as to what other information is needed that will appropriately inform interpretation and synthesis.

The Demands on the User for Interpreting Scores on Number Series Items

A set of released items from TIMSS (The Third International Mathematics and Science Study) was administered to 279 students in seventh grade in a southwestern school district. One number series two-part item presented a diagram of three different sized but congruent right triangles. Small triangles like the first in the series were embedded in the second and third triangle of the series. Part One of the item was to determine the number of triangles in each of the three triangles. It turns out to be 1, 4, and 9. Eighty-two percent of the students responded correctly to this simple task of counting the number of small triangles in the two larger triangles. The second part of the item asked the students to determine how many triangles would be needed for triangle 8 if the series were extended to the 8th figure in the series. The southwest students had a 3% correct response rate and 49% no response rate compared with the international 18% correct response. Analysis of student responses revealed that many of them tried to draw triangles up to the eighth in the series and got bogged down in graphic representation. These students “knew” how to complete some number series since a high percentage of the population correctly identified “t = 7 x 8” as the correct “number sentence” to determine the eighth number in the pattern of 7, 14, 21, 28 on the state assessment. Further information on context makes interpretation more complex. Teachers were trained in mathematics problem solving instruction with emphasis on students paired-off to justify their problem solving to each other with teacher monitoring. A survey revealed most teachers reported conducting paired problem solving with justification of responses at least three times a week. This was confirmed by observation and interview of a non-random sample of teachers. But there were differences in the quality of the justification with some teachers encouraging “assistance” in doing the problem and shortcutting the justification process in part due to difficulty in managing small group work. Thus interpretation of this one number series item makes demands on the user with implications for validity studies of student thought processes, the context of instruction, and ultimately interpretation of findings. Evaluators must go beyond the test to look for impacts such as transfer (or lack of it) in this case and the use of tracing of cognitive processes added to the assessment of outcomes.

Contributions to Reasoning from Test Scores Made by Reviewers of Educational Tests and Assessments

One way of examining the contributions of reviews of student achievement test batteries to appropriate reasoning from test scores is to examine the reviews in the Buros Mental Measurement Yearbooks (MMYs) produced by the Buros Institute of Mental Measurements at the University of Nebraska and distributed by the University of Nebraska Press. The reason this is a good source for examining current practice is that reviewers are chosen with the expectation that they will write with MMYs objectives in their "reviewer's guide" in mind: "to provide test users with carefully prepared appraisals of test materials for their guidance in selecting tests; to stimulate progress toward higher professional standards of test construction by commending good work, by censuring poor work, and by suggesting improvements; and to impel test authors and publishers to present more detailed information on the construction, reliability, norms, uses, and possible misuses of their tests" (Buros Mental Measurements Institute, undated). The limitation in this source is that today state departments and districts collaborate in development of achievement tests and those assessments are not subject to this type of review (Wise, 2008). Nevertheless state and district tests are usually adaptations, with some local tailoring of items, of publishers' tests that are reviewed. Also, while it is true that local districts do contract with experts to review their tests, these are rarely available to the general public or other professionals.

The most recent study of the Buros reviews is by Cizek, Rosenberg, and Koons (2008). They analyzed reviews of all 283 tests in the *Sixteeenth Mental Measurements Yearbook* (16th MMY) (Spies and Plake, 2005) focusing on conformity to modern validity theory, sources of validity evidence typically reported, differences across kinds of tests, and validity factors considered most important. Achievement tests (the focus of the present article) constituted 19% of the 283 tests in the ten types of tests included in the 16th MMY and their review. Most relevant to the current paper is that the lowest percentages of mention of validity evidence for achievement tests were: evidence of predictive validity (test-criterion relations) 18.5%; evidence of test consequences, 0.0%; and evidence of cognitive response processes (3.7%). It is important to note here that meeting the criterion for "sources" of evidence in Cizek and others (2008) is not necessarily an indication of "quality" of evidence of validity. For example, evidence of construct validity may involve convergent evidence of a strong correlation of the achievement test with a similar achievement test say in reading comprehension and divergent evidence in the form of a lower or moderate correlation with a problem solving test in mathematics. However that does not mean that there is evidence of the reading comprehension construct (or domain of desired accomplishment) underlying the test, such as perhaps "students putting the meaning of a paragraph in their own words" nor of experimental evidence. Cizek and other (2008) review, however, does show that certain aspects of validity are being well described in test reviews. These areas include representativeness of content to national trends; development of norms for comparison of local performance with national normative groups; attempts to control for bias and construct irrelevant variables; and provision of manuals or guides to administration, interpretation and use.

The challenge to the profession comes from an apparently rather low rate of gathering or presenting evidence on consequences of testing, response processes of test takers, and assessment that integrates validity evidence from multiple sources. Caution to users by reviewers are often mentioned, but not for all key issues. For example, reviewers may mention how norms were developed and not mention that norms may need to be developed for the local intended use. Reviewers may report validity evidence, but if it does not include evidence appropriate to planned uses, the user is not always reminded of his responsibility for that evidence. Reviewers may refer to publisher guides to interpretation but not note that the guides have not been subject to evaluation of effectiveness.

What becomes clear from this line of analysis is the opportunity for university researchers, psychometricians, teachers of educational measurement, school administrators, and evaluators to appropriately influence practice in the interpretation of tests, synthesis of multiple lines of evidence, and evaluating those processes as meta-evaluations. That opportunity might well be taken up as a professional responsibility.

Implications for a Line of Practical Action by the Evaluation Profession

Implications for evaluators have been noted throughout the article to open up a dialogue on bringing to practice the contributions of evaluation theorists to understanding interpretation, synthesis, and argument or reasoning from data. In part it is a matter of exploring the "why" of the outcomes as suggested decades ago (Hastings, 1966). A rather thorough survey of evaluator competencies (Dewey and others, 2008) while not claimed as definitive reveals no specific attention to skills in integrating information from multiple sources, and constructing a validity argument for an interpretation or synthesis from that chain of evidence. However, evaluation theorists have already articulated relevant perspectives and strategies. Most pointedly and broadly, the work of Scriven (1991, 2007, 2009) and Scriven elaborated by Davidson (2005) on evaluation synthesis and meta-evaluation, Schwandt (1998, 2008a, 2008b, 1998) on practical knowledge and evaluation as reasoning, and House (1980, 1995) on evaluation as argument have addressed the complexity of integrating the best evidence with clinical expertise and values embedded in means and goals of interpretation. But what is needed is deliberation on bringing these perspectives into common practice. These perspectives, taken together, point to the need to capture a broad range of relevant data to make sense of the why of test scores (outcomes) and integrating that into an assessment—an interpretation, synthesis, and argument.

The body of the article, is intended to make more transparent issues for deliberation around bringing current theory into practice. We conclude this section with a suggestion for development of exemplars to guide practice in the form of casebooks. Robert Linn (2006) proposed as a supplement to testing standards "... to mount an effort to create a casebook that would provide realistic examples of applications of the standards to tests and testing programs ... showing how the principles might apply in specific situations". In the context of evaluation, supplements to the evaluation standards and guiding principles are needed. This is particularly so for gathering information to allow understanding of the why of the outcomes and for the process of interpreting findings.

Casebooks demonstrating interpretive processes in different contexts of evaluation practice would serve the profession well. The need for this kind of guidance is indirectly supported by the Guiding Principles for Evaluators (AEA, 2007) and The Program Evaluation Standards (Yarborough et al., 2011). What is made clear in these sources and responses from the field to these evaluation guidelines is that principles and standards do not imply an algorithmic use. Expertise is required for applying standards, using information at hand or accessible and responding to the demands of context. This complexity in use of evaluation standards has been well illustrated by the profession (e.g. see Kirkhart, 2008). However, exemplars of expertise in casebooks for different contexts can do much to move the profession in the direction of current theories of practice in reasoning from information to interpretation. In addition to the sources noted above one might add the work of theorists on development of outcome measures to enhance interpretation and argument. Two useful sources are Funnel and Rogers (2011, p. 179) on generation of outcome chains and Patton (2008, pp. 300-305) on alternative ways of focusing an evaluation and the kinds of questions that come out of that approach.

References

AERA, APA, NCME (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Evaluation Association (AEA). (2007). *Guiding Principles for Evaluators*. Retrieved 14 March 2011, http://www.eval.org/publications/aea06.GPBrochure.pdf

Berliner, D. C. (2009). Our impoverished view of educational reform. *Teachers College Record, 108* (66),949-996.

Brickell, H.M. (1976, 2011). Needed: Instruments as good as our eyes. *Journal of MultiDisciplinary Evaluation, 7*(15), 171-179.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, MA: Harvard University Press.

Buros Mental Measurements Institute (undated). *Reviewers Guide*. Retrieved 31 May 2011. http://www.unl.edu/buros/bimm/html/suggestions.html

Cizek, G. J., Rosenberg, S. L., and Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397-412.

Davidson, E. J. (2005). *Evaluation methodology basics*. Thousand Oaks, CA: Sage.

Della-Piana, G. M. (2008). Enduring issues in educational assessment. *Phi Delta Kappan, 89*(8), 590-592.

Dewey, J. D., Montrosse, B. E., Schroter, D. C., Sulllins, C. D., & Mattoz II, J. R. (2008). Evaluator competencies: What’s taught and what’s sought. *American Journal of Evaluation, 29*, 268-287.

Fu, A. C., Raizen, S. A., and Shavelson, R. J. (2009). The nation’s report card: A vision of large-scale science assessment. *Science, 326*, 1637-1638.

Funnell, S. C. & Rogers, P. J. (2011). *Purposeful program theory: Effective use of theories of change and logic models*. San Francisco, CA: Jossey-Bass.

Herszenhorn, D. M. (May 5, 2006). As Test-Taking Grows, Test-Makers Grow Rarer. *New York Times*. Retrieved 16 March 2011. http://www.nytimes.com/2006/05/05/education/05testers.html?_r=1&scp=1&sq=As+test-taking+grows+&st=nyt/

Hastings, J. T. (1966). Curriculum evaluation: The why of the outcomes.

Journal of Educational Measurement, 3(1), 27-32.

House, E. R. (1980). *Evaluating with validity*. Beverly Hills, CA: Sage.

House, E. R. (1995). Putting things together coherently: Logic and justice. In D. Fournier (Ed.), *Reasoning in evaluation: Inferential links and leaps. New Directions for Evaluation, 68*. San Francisco: Jossey-Bass.

Joint Committee on Standards for Educational Evaluation (2001). *The Student Evaluation Standards: How to Improve Evaluations of Students*. Thousand Oaks, CA: Corwin Press.

Kirkhart, K.E. (2008). Commentary: Consumers, culture, and validity. In M. Morris (Ed.). *Evaluation ethics for best practice: Cases and commentaries* (pp. 31-53). New York: Guilford.

Linn, R. L. (2006). Following the standards: Is it time for another revision? *Educational Measurement: Issues and Practice, 25*(3), 54-56.

Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17*(2), 28-30.

Lissitz, R. W. (2009) (ed.). *The concept of validity: Revisions, new direction, and applications*. Charlotte, NC: Information Age Publishing.

Lohman, D.F. & Nichols, P. (2006). Meeting the NRC panel's recommendations. *Educational Measurement: Issues and Practice, 25*(4), 58-64.

Madaus, G., Russell, M. & Higgins, J. (2009). *The paradoxes of high stakes testing*. Charlotte, NC: Information Age Publishing.

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.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Pellegrino, J. W., Chudowsky, N., and Glaser, R. (Eds.) (2001). *Knowing what students know: The science and design of student assessment*. Washington, DC: National Academy Press.

Schwandt, T. A. (2008a). Educating for intelligent belief in evaluation. *American Journal of Evaluation. 29*(2), 139-150.

Schwandt, T. A. (2008b). The relevance of practical knowledge traditions to evaluation practice. In N. L.Smith & P. R. Brandon (Eds.). *Fundamental issues in evaluation* (pp. 29-40). NY: Guilford.

Schwandt, T. A. (1998). The interpretive review of educational matters: Is there any other kind? *Review of Educational Research. 68*(4), 409-412.

Scriven, M. (2009). Meta-evaluation revisited. *Journal of MultliDisciplinary Evaluation, 6*(11), iii-viii.

Scriven, M. (2007). *Key evaluation checklist*. Retrieved 17 March 2011 from http://www.wmich.edu/evalctr/checklists/metaevaluation/

Scriven, M.(1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.

Stufflebeam, D. (2011). *Meta-evaluation checklists*. Retrieved 17 March 2011 from http://www.wmich.edu/evalctr/checklists/checklistmenu.html

U.S. Department of Education. *Race to the top assessment funding*. Retrieved 11March, 2011.

http://www2.ed.gov/programs/racetothetop-assessment/index.html.

Wise, L. L. (2006). Encouraging and supporting compliance with standards for educational tests. *Educational Measurement: Issues and Practice*, *25*(3), 27-34.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., and Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage