Priority Validity Considerations for Formative and Accountability Assessment[1]

Eva L. Baker
University of California, Los Angeles (UCLA). National Center for Research on Evaluation, Standards, and Student Testing (CRESST). California, United States of America

Abstract

This paper focuses on validity of testing systems (Baker, 2007a) used in schools intended to improve instruction, either through formative or accountability-based assessments. The argument put forward supports the need for studies of the instructional sensitivity of tests to assure that they are actually measuring school effects. The research agenda suggested by these concerns is extensive and essential. Eva L. Baker presents CRESST (Center for Research on Evaluation, Standards and Student Testing, University of California, Los Angeles) research jobs related to instructional validity and main issues related to Criterion Referenced Test (CRT). Finally, in order to conduct such studies, better and more scalable measures of classroom instruction are needed.

Key Words: Assessment and accountability, validity studies, formative assessment, instructional validity, educational assessment.

Background

Accountability is a function desired world-wide, and it has risen to enormous importance in the area of educational policy. Accountability, in simple terms, means holding an individual, group, agency, or government responsible for its actions and their consequences as they attempt to meet specified goals. In the USA, educational accountability (Baker, in press; Baker, Goldschmidt, Martinez & Swigert, 2002) has mostly focused on the schools, with an emphasis on the performance of students in limited curriculum domains (mathematics, English reading and writing and science) at the elementary and secondary school levels. In other countries, accountability may be focused on a wider range of subject matters and on the performance of the individual student. In some settings, accountability and the consequences for inadequate student or school performance are broadly shared among sectors, and the sanctions encompass students, teachers, administrators, policymakers, and politicians. This article will, as you might suppose, focus principally on the experience in the USA and will be offered from the perspective of a person working in the research and development arena.

A review of the progress of US students since the most recent push for performance began (around 1992) is not encouraging US students currently perform near the middle on international comparisons (Organisation for Economic Co-operation and Development, 2005 [PISA 2003]). More serious, of course, is the US achievement gaps between minority and majority students, where progress is

([1]) The work reported herein was supported under the National Research and Development Centers, PR/Award Number R305A050004, as administered by the US Department of Education's Institute of Education Sciences (IES). The findings and opinions expressed in this report do not necessarily reflect the positions or policies of the National Research and Development Centers or the US Department of Education's Institute of Education Sciences (IES).

being made, but not at a rate sufficient to change the overall performance of groups in any reasonable time (Baker, Griffin & Choi, 2008). In this respect, looking at equity in countries with immigrant populations the US is not alone in its struggles (Organisation for Economic Co-operation and Development, 2005).

While *No Child Left Behind* (2002), the name of the legislation focusing accountability attention on US schools and students, has been implemented with a plan and speed that, while understandable, has encouraged states (because of our decentralized educational system) to select examinations that may not be the best measures of school effectiveness and student learning. What is needed in an accountability system are both credibility and validity of the measures used to mark progress to the goals, valid and achievable standards, and, of course, a rational plan for how to use the available financial and human resources to attain desired outcomes.

Validity is a well known attribute attached to tests, conveying in general the idea of their appropriateness, technical quality, and relevance to goals. Of course, validity has more complex interpretations, the most far-reaching of which was put forth by Messick (1989) where he focused on the quality of the decisions made from results rather than features of the test itself. As the uses of tests have expanded and their nominal purposes have multiplied, so has the complexity and sophistication required of validity studies. Yet, because of the schedule of implementation, very few validity studies are fielded in the US before the assessments are applied as measures of effectiveness, and most validity studies never address the full range of purposes articulated for examinations. Instead of focusing on the entire set of purposes for which accountability tests (and formative assessments) may be directed, we will describe one cluster of purposes that is challenging but difficult to address. Perhaps by pooling international expertise, we may be able to make progress in these areas of validity.

What are the purposes that are essential for tests to exhibit if they are to serve as high quality proxies for educational effectiveness? In an article published about 15 years ago, my colleagues and I enumerated criteria that assessments and tests should exhibit if they are to be useful in the US accountability scene (Linn, Baker & Dunbar, 1991; Baker, O'Neil & Linn, 1993). The criteria relate in part to the goals or standards that the assessments were to sample in light of the information to be gleaned from their results, either annually or over a longer time interval. Included in this list were the characteristics of cognitive complexity (that is, asking students to do significant processing of content and generating or analyzing using strategies that required multiple steps). A second feature related to the richness of the content in which the cognition was embedded. This criterion extended to the significance and accuracy of the content as measured. Both of these criteria have implications for how performance was to be judged, for example, whether scoring criteria or rubrics illustrated the levels of cognition and content knowledge desired. A third criterion was fairness, to assure that the assessment provided an even-handed opportunity for students from different backgrounds or experiences in the same instructional settings to demonstrate competence. Inherent in fairness is the avoidance of construct-irrelevant variance, the extent to which prior knowledge, test formats, or content systematically advantage particular groups of students. Fairness also applied to the access and capacity of students with disabilities to demonstrate competency on the examination. A broader interpretation, one to which we will return, involved the extent to which students had been given reasonable opportunity to learn the material measured on the test (see *Raising Standards for American Education*, National Council on Education Standards and Testing, 1992). A fourth area for validity criteria was the extent to which the test, or testing system more accurately, provided opportunity for students to demonstrate transfer of learning and generalizability of skills, strategies and knowledge. These features are essential to ensure that student-demonstrated procedures or problem-solving (Vendlinski, Baker & Niemi, 2008) represent their serious understanding and expression rather than as a product of rote processes. While all of the forgoing validity features are important, the rest of this article focuses on the one cluster that is required if assessments or examinations are to be used as adequate bases for making inferences about educational effectiveness. The major validity criterion here is *instructional sensitivity.* Instructional sensitivity is a key component of any comprehensive validity argument in support of tests used in educational accountability systems. Sensitivity to instruction (Baker, 2008) can be inferred from evidence that scores on a test (examination

or assessment) are differentially affected by high-quality instruction, ideally given over significant portions of the school year. The 'high-quality' descriptor is intended to differentiate complex instruction from drill or direct practice of highly similar test items that affects test scores and little else. This tautological tinge has earlier been explored in research related to aptitude or trait-treatment interactions. (See, for example, Berliner & Cahen, 1973; Cronbach & Snow, 1977; Tobias, 1976. One conclusion growing from research in this area was that outcome measures required specificity if they were to be sensitive to treatments.)

This article considers why instructional sensitivity should be documented, the sources for recommendations for assessment design to increase such sensitivity, and resulting desirable, conjoint features of tests and instruction. Instructional sensitivity comes into play when aspects of the following situation hold: (1) there is great variation in the implementation of an explicit *curriculum* or instructional plan; (2) there is substantial difference in the preparation and resulting quality of teachers; (3) teachers are specifically given flexibility to adapt the *curriculum*; and (4) there is no explicit *curricula* or plan of studies other than an enumeration of standards to be achieved (true of many US settings). Instructional sensitivity will be addressed, first with an argument for its importance, then a brief review of past strategies for studying instructional sensitivity in schools, and last, a brief consideration of recommendations for future validity studies.

Why Instructional Sensitivity Matters?

When measures of achievement serve as the principal determiner of judgments about educational effectiveness in an accountability system, some level of evidence is required that shows a defensible relationship between scores on a test (assessment or examination) and the quality of instruction or other in-school educational experiences. Unless there is a convincing causal link between the two, the use of such tests as dependent measures hovers between questionable and unwarranted. The direction of the linkage is not that the tests are designed to match instruction, but rather that the goals stimulate appropriate instruction activity and the consequent learning sampled on the test. This mapping of goals to instruction to assessment is what alignment is supposed to be about (Baker, 2005), but it is difficult to fully implement for reasons discussed later.

Validity evidence is also expected for each discrete test purpose (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999). For example, the logic underlying many accountability systems implies more than the purpose to detect differential effectiveness of students' instructional experiences. In addition, there are purposes related to system improvement over time, as exhibited in the patterns of growth targets specified for Adequate Yearly Progress (AYP) in *No Child Left Behind* or any other approach permitting the incremental attainment of pre-specified performance levels. For such improvement cycles to occur, test results should aid in the selection of instructional materials, the diagnosis and teaching of the current crop of students (if timing of results permits) or, more likely, of future cohorts' students (as the system improves over time). So it is sensible to determine as part of validity studies, the subordinate but necessary pieces of evidence showing whether teachers can draw inferences from student results, apply them to students' performance shortfalls, and develop or apply a range of alternative instructional strategies for students with gaps, misconceptions, or other inadequacies. These behaviors are essential to the theory of action underlying all accountability systems that intend to measure the deliberate effects of teaching and schooling (Baker & Linn, 2004). The logic is also key to the concepts of formative assessment (Black & Wiliam, 2003; Pellegrino, Chudowsky & Glaser, 2001). Of course, there is a critical intermediate link in all of these purposes, to wit, that teaching actually results in learning, however measured. Let us consider instructional sensitivity and its derivation from the design of earlier assessment

systems more intimately linked to instruction and improvement. Then, as a contrast, we will review the design and purposes of modal, current standard-based systems.

The Derivation of Instructionally Sensitive Measures

In the 1960s and 1970s, stimulated by an emphasis on self-instructional systems, the concept emerged of Criterion-Referenced Tests (CRT) or its elaboration, Domain-Referenced Tests (DRT). These approaches to measurement were treated as novel and advocated by leading scholars of the time, many of whom had worked on the development of programmed instruction or other integrated instructional systems before they took up the topic of measurement. CRTs were contrasted with norm-referenced examinations – designs intended to result in normal distributions of results in order to permit ordered comparisons of examinees (Popham &Husek, 1969). CRT as a term was coined by Glaser (1963), but represented in earlier work (Lumsdaine & Glaser, 1960). The elaboration of CRT design by scholars in the math and science areas linked it to set theory (Hively, Patterson & Page, 1968) who both invented and explicated the properties and potential benefits of domain-referenced testing.

CRT Design Issues

Chief among the attributes of criterion-referenced tests was at its starting point of design: a well-specified domain of content and intellectual (now cognitively-oriented) skills. Coming from a behaviorist view of learning, design of these front-ends was very specific, and over time moved away from simply detailed enumeration to include more developed and rationalized cognitive demands. Actually, the level of specificity of these measures was much like that needed at the time to create workable, replicable, instructional sequences. Instead of beginning with a general construct and test specifications that distributed items in one or more formats across a wide range of topics, broadly but lightly sampled, the new alternative called for well-defined content and skill boundaries including content domain, problem types, formats, and scoring rules that would delimit membership of tasks or items in the set (note that the domain specification did not imply a false uniformity, for difficulty levels were assumed to vary within a domain as a function of features such as conceptual difficulty, confusion with other similar ideas and concepts, the embedded context of the tasks and, of course, the level of criteria set for different degrees of proficiency). As the ideas emerged, so did the probabilistic notion of item membership (e. g., fuzzy sets), to characterize the legitimacy of membership of task or item samples within the explicit features of the domain. It was expected that an extensive and carefully constructed sample of items/tasks would deeply and directly measure performance either within the total domain, or for its particular sub-parts.

These domain boundaries or rules were intended to have instructional power as well because they could be made transparent (in clearer language and with supporting examples) and directly shared with teachers expected to manage instruction. Teachers who generated or applied instruction which exhibited these explicated domain features were expected to produce predictable performance gains on the CRTs, much like instructional systems were expected to be effective after a cycle or two of revision. Because the impetus of CRTs came from the design of systematic instructional systems, the expectation that students with less than satisfactory performance would get additional chances, with more refined instruction, to learn and then to demonstrate their accomplishments as improved test performance. Such an iterative view has characterized training for a number of years. In my view, this notion of defining the universe

from which test tasks could be fairly drawn (Hively et al., 1968) was the most powerful part of the notion of criterion-referenced testing.

CRT Reporting

Another, more salient and easy to implement feature of CRT, focused on reporting. CRT results would be reported differently from commonly available commercial tests. CRTs were intended to convey some notion of competence or mastery with respect to a (perhaps, arbitrary) corpus of knowledge and skills rather than point to a particular spot on the continuum of a broad ability construct. Thus, CRT reports included at the outset crude metrics like percent correct or percent achieving some largely arbitrary target (90% will achieve 90%). These numbers appeared to give a more definitive picture of students' acquisition of targeted competencies when contrasted with transformed normative values reflecting the student's place in the distribution of examinees, but were easily subject to manipulation. Setting cut scores in order to divide the group into various levels of competence, itself generated a cottage industry on how to develop cut-scores and attendant verbal descriptions to characterize student results. Presumably these levels were to be validated as well. It was the reporting part of CRT, however, that captured the attention of test developers and users. It was relatively easy to convert any record of performance into frequency distributions within a particular score range or ranges, set by performance standards or cut scores. In fact, the reporting protocols swamped the design requirements, setting us up for current large-scale assessment practices. We now have standards-based assessments in the US and worldwide, and they share only the nominal features of the CRT model.

The Situation Now–Standards-Based Assessments

Standards-based testing systems are often portrayed and understood as if they were CRTs, and that is certainly the impression given by the notion of ‘standards-based’, at least as practiced in much of the United States. These systems, in fact, use standards as the source of their design, whose verbal descriptions circumscribe only loosely the tasks or items within. Rather than using the design rules for item membership, standards-based tests have adopted the reporting characteristic of CRTs. Most obviously, they use criterion points as cut scores among categories such as Below Basic, etc. These are set by complex approaches relating to items whose intent may or may not survive subsequent scaling procedures. On the design front, the details of the assessments are less well understood. They rarely invoke actual construct validity studies that might be appropriate to their design structure. It is a case of having the ‘name but not the game’ in American slang.

Instructional alignment to standards and assessments under these conditions presents a challenging problem. Standards, it is thought, should guide instruction and should fashion classroom focus. Tests were to be merely indicators of acquisition and application of skills and knowledge. As it is said in Los Angeles, California, ‘yeah, right’ (a sarcastic double positive that means a negative).

A range of procedures (see, for example, Herman & Webb, 2007) can be found to decide how much alignment has been achieved (Baker, 2005). For the most part, these approaches look at the referencing of content in standards to nominally relevant test items. Although there are some efforts at establishing the depth of measurement (Webb, 1999), it is by no means a universal criterion, and thus many statewide tests may have relatively few items, or as few as one or a fraction of one, addressing a standard or one of its component standards. The reason for the decision to use survey measures include

the time allotted for testing, the speed of implementation, and cost–all major barriers to better task sampling plans.

However, when standards are not adequately measured (perhaps because of their number, or time and cost constraints), what should a teacher facing sanctions do? It is far wiser for that teacher to avoid sanctions by teaching the topics that occur on tests with some frequency. They can get this information from test inspection or by personally or by surrogate inspecting blueprints to determine what is actually included on the test. The result is well known, that practice on test content has become the norm in many schools, and in particular, for those at risk for sanctions. The consequences of this approach are many: efforts to teach the standards are forgone in favor of teaching tested content; coherence and cumulative knowledge is lost; students who are underperforming have little systematic attention to subtask learning that might support multiple standards and future performance (in other words, skills are in silos and separated from one another). The acquisitions of procedures or of procedural tricks may be easier to teach than difficult concepts in rich applications.

This is not a big deal in the policy realm where many believe that the details of particular tests do not matter as much as their titles, and nuances, such as depth of sampling, may be imagined to be a technical but not essential feature. When measures are strongly correlated, it has been publically opined by policymakers, they are exchangeable for one another. So, even with standards-based assessments, many of the traditional approaches to task or item development and sampling are used rather blithely.

Multiple Purposes and Instructional Sensitivity

As the requirements for accountability measures evolved, it became ever clearer that tests were expected to serve different purposes, some noted above. Most experts in psychometrics will quickly point out that each test purpose deserves its own test in order to optimize the quality of the ensuing decision to be made, whether admission, placement, or system monitoring. I do not totally agree, but I am clear that tests created for one purpose can rarely be retrofitted to serve a different purpose. Of course, the retrofit process was the only way that accountability tests could be created quickly and cheaply. However, if such tests were principally designed using general specifications rather than well-specified domains or ontologies, how can the concept of sensitivity of instruction be addressed?

High Quality Instruction?

The question is a simple one. How do you know that measures nominally assessing desired outcomes actually are sensitive to high-quality instruction? If we have sketched out features on the measures side that might make them sensitive to instruction, how should quality instruction be characterized? First of all, and as usual, instruction must relate substantively to the goals to addresses. It should represent the key cognitive and content elements arrayed in sequences designed to assure prerequisites. In the best case, the instruction should provide principle-based knowledge and strategies supporting relevant and prerequisite knowledge, so that students understand why they are using particular approaches and where they fit in the domain or sequence of instruction. To the extent possible, instruction should be guided by findings from research on learning that attempt to minimize memory or cognitive load (Sweller, 1999), to support schema development (Chi, Glaser & Farr, 1988), and to provide relevant feedback. To support transfer, instruction should involve a range of formats for tests as well as other applied and integrated approaches to display accomplishments. These include feedback attached to appropriate domain practice, the

understanding of purpose (motivation), individual differentiation, and a graduated sequence of learning all wrapped in the active engagement of students (see Popham & Baker, 1970, for an early treatment of these concepts in a teacher preparation context). Such applications of pedagogical principles are feasible and reasonably effective. Yet, the degree of intensity, mix, and frequency of these and other principles of instruction, scaffolding, and levels of engagement, for instance, might well suggest refined validity studies. If we can agree on the 'quality' part of instruction, we can invert our analyses and then see what kind of test responds best.

Presumptive Conditions for Instructional Sensitivity: An Interim Summary

To study instructional sensitivity, there are some pre-cursive questions that should be answered in the affirmative, so at least there is a higher if not 'high' quality treatment to be included. These questions are clear and need little elaboration:

- Do test results communicate operational domains for instructional action? Are they understood by teachers?
- Is there expertise and time for results-based teaching?
- For attention to differences?
- Can teachers augment or investigate where help is needed (high-quality formative assessment)?

How Do We Know? Evidence for and against Instructional Sensitivity

Referenced earlier was a list of validity criteria for assessments (Baker, O'Neil & Linn, 1993). The validity criteria specified in that paper provided much of the intellectual glue of the CRESST R&D agenda for the next two decades. Key questions that could be answered by inspecting state data surely give us a gross estimate of the probability of finding instructional sensitivity:

- What is the relationship of the amount of test practice to the total functional instruction time and how is it connected to test results?
- Are there differences in the exposure rates to this practice by subgroups?
- Are there instructional treatments focused on rapid acquisition of prerequisites?
- Is performance on transfer tasks assessed?
- Is there evidence from the data that individual learning is sustained and cumulative?

One guess is that present levels of evidence either are blank or only modestly compelling in many of our states.

CRESST Experiences with the Instructional Sensitivity Issue

As part of our research, we promised to address the validity criteria that were later to appear in the Linn et al. paper, and used the criteria as glue to link studies of assessment design and validation. CRESST undertook the development of yet another incarnation of CRT in 1990 called Model-Based Assessments (Baker, 1997, 2007b) earlier termed Cognitively-Sensitive Assessments. It was our attempt to develop assessments that would be sensitive to instruction, that is, to reflect a complex domain whose features could be used to guide teachers' or others' instructional planning. In addition, we had questions about what attributes of the assessments could be interpreted as domain-independent, that is, useful across different topics and subject matters, which were focused on approaches drawn from the literature on learning a particular domain (see *Knowing What Students Know*, Pellegrino, Chudowsky & Glaser, 2001, for a discussion), and whether the development of such measures could focus first on cognitive demands that might be domain-independent, and thus reduce the cost of developing subsequent high-quality measures. We had a number of other bright ideas, for instance, that teachers would use automated (smart) tools to help them design classroom testing, something that we found was largely a non-starter, as others had learned (Vendlinski & Stevens, 2000) even when teachers liked the process, knew how to use procedures, and had the correct technological support. The wave of accountability, even before NCLB, leached away teachers' time, their interest, and confidence in developing their own relevant classroom measures.

Measuring Classroom Practice

CRESST with a large group of the educational research community have invested in the measurement of classroom practice, or opportunity to learn (Aguirre-Muñoz, Boscardin, Jones, Park, Chinen, Shin, Lee, Amabisca & Benner, 2006; Herman & Abedi, 2004; Herman & Klein, 1997; Yoon & Resnick, 1998). Approaches used include observations, self-reports of knowledge and pedagogy (Baker, Niemi, Herl, Aguirre-Muñoz, Staley, Linn & Rogosa 1996), observations (Boscardin, Aguirre-Muñoz, Chinen, Leon & Shin, 2004), student assignments and student work (Aschbacher, 1999; Clare & Aschbacher, 2001; Matsumura, Garnier, Pascal & Valdes, 2002; Matsumura & Pascal, 2003; Matsumura, Slater, Wolf, Crosson, Levison, Peterson, Resnick & Junker, 2006), and interpretation by teachers of results and instructional plans (Herman & Baker, 2003).

These studies at CRESST were generally conducted in order to determine how our assessments scaled up, or in other words, the degree to which fidelity conflicted with classroom reality (Baker, Niemi, Herl, Aguirre-Muñoz, Staley, Linn & Rogosa 1996). We have also conducted process-product studies using existing external measures (Goldschmidt, Martinez, Niemi & Baker, 2007) (on performance of CRESST test and the high school exit exam), often in the context of evaluating an intervention. In these studies we search for relationships among self-report, observation, or artifacts (Matsumura, Garnier, Pascal & Valdes, 2002; Stecher, Borko, Kuffner, Martinez, Arnold, Barnes, Creighton & Gilbert, 2007). All such studies, while interesting when there are low relationships, almost never produce convincing data because of the confounding of teacher quality with student assignment, a fact pointed out by Rogosa's odds-ratio work (Rogosa, 1999a,b,c).

References

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION (2008). *Unleashing the power of formative assessment: A strategy for integrating cognitive research, assessment, and instruction*. Session 13.028 at the annual meeting of the American Educational Research Association, New York.

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.

AGUIRRE-MUÑOZ, Z., BOSCARDIN, C. K., JONES, B., PARK, J.-E., CHINEN, M., SHIN, H. S., LEE, J., AMABISCA, A. A. & BENNER, A. (2006). *Opportunity to learn measures* (CSE Rep. 678). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

ASCHBACHER, P. R. (1999). *Developing indicators of classroom practice to monitor and support school reform* (CSE Rep. 513). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

BAKER, E. L. (in press). Learning and assessment in an accountability context. In K. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability*. Mahwah, NJ: Erlbaum.

— (1997). Model-based performance assessment. *Theory Into Practice*, 36, 247-254.

— (2005). *Aligning curriculum, standards, and assessments: Fulfilling the promise of school reform*. In C. A. DWYER (Ed.), Measurement and research in the accountability era (315-335). Mahwah, NJ: Erlbaum.

— (2007). Model-based assessments to support learning and accountability: The evolution of CRESST's research on multiple-purpose measures. *Educational Assessment* (Special Issue), 12(3&4), 179-194.

— (2008). Empirically determining the instructional sensitivity of an accountability test. Paper presented at the annual meeting of the American Educational Research Association. In W. James Popham, *Empirically Determining the Instructional Sensitivity of an Accountability Test: Alternative Approaches*. New York.

BAKER, E., GOLDSCHMIDT, P., MARTÍNEZ, F. & SWIGERT, S. (February, 2002). *In search of school quality and accountability: Moving beyond the California Academic Performance Index (API)* (Deliverable to OERI, Contract No. R305B6002). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

BAKER, E. L., GRIFFIN, N. C. & CHOI, K. (in press). *The achievement gap in California: Context, status, and approaches for improvement*. Paper prepared for the California Department of Education, P-16 Closing the Gap Research Council «Connecting the Dots and Closing the Gap». Davis, CA: University of California, Center for Applied Policy in Education (CAP-Ed).

BAKER, E. L. & LINN, R. L. (2004). Validity issues for accountability systems. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 47-72). New York: Teachers College Press.

BAKER, E. L., NIEMI, D., HERL, H., AGUIRRE-MUÑOZ, Z., STALEY, L., LINN, R. L. & ROGOSA, D. (1996). *Report on the content area performance assessments* (CAPA): A collaboration among the Hawaii Department of Education, the Center for Research on Evaluation, Standards, and Student Testing (CRESST) and the teachers and children of Hawaii (Final Deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

BAKER, E. L., O'NEIL, H. F., JR. & LINN, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist*, 48, 1210-1218.

BAKER, E. L., PHELAN, J., CHOI, K., NIEMI, D., VENDLINSKI, T., GRIFFIN, N., HERMAN, J. L. & HOWARD, K. (in progress). *Design and validation of POWERSOURCE© assessments and instructional materials*. Los Angeles, University of California, National Center for Research on Evaluation, Standards, and Student Testing.

BERLINER, D. C. & CAHEN, L. S. (1973). Trait-treatment interaction and learning. *Review of Research in Education*, 1, 58-94.

BLACK, P. & WILIAM, D. (2003). In praise of educational research: formative assessment. *British Educational Research Journal*, 29, 623–637.

BOSCARDIN, C. K., AGUIRRE-MUÑOZ, Z., CHINEN, M., LEON, S. & SHIN, H. S. (2004). *Consequences and validity of performance assessment for English learners: Assessing opportunity to learn (OTL) in grad 6 language arts* (CSE Rep. 635). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

CHI, M. T. H., GLASER, R. & FARR, M. (Eds.). (1988). *The nature of expertise*. Hillsdale, NJ: Erlbaum.

CLARE, L. & ASCHBACHER, P. R. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. *Educational Assessment*, 7, 39-59.

CRONBACH, L. J. & SNOW, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

GLASER, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519-521.

GOLDSCHMIDT, P., MARTINEZ, J. F., NIEMI, D. & BAKER, E. L. (2007). Relationships among measures as empirical evidence of validity: Incorporating multiple indicators of achievement and school context. *Educational Assessment* (Special Issue), 12(3&4), 239-266.

HERMAN, J. L. & ABEDI, J. (2004). *Issues in assessing English language learners' opportunity to learn mathematics* (CSE Rep. 633). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

HERMAN, J. L. & BAKER, E. L. (2003). *The Los Angeles Annenberg Metropolitan Project: Evaluation findings* (CSE Tech. Rep. No. 591). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

HERMAN, J. L. & KLEIN, D. C. D. (1997). *Assessing opportunity to learn: A California example* (CSE Rep. 453). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

HERMAN, J. L. & WEBB, N. M. (2007). Alignment methodologies. *Applied Measurement in Education*, 20, 1-5.

HIVELY, W., PATTERSON, H. L. & PAGE, S. H. (1968). A «universe-defined» system of arithmetic achievement tests. *Journal of Educational Measurement*, 5, 275-290.

LINN, R. L., BAKER, E. L. & DUNBAR, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20, 15-21. (ERIC Document Reproduction Service No. EJ 436 999)

LUMSDAINE, A. A. & GLASER, R. (Eds.). (1960). *Teaching machines and programmed learning: A source book*. Washington, DC: National Education Association of the United States.

MATSUMURA, L. C., GARNIER, H. E., PASCAL, J. & VALDES, R. (2002). *Measuring instructional quality in accountability systems: Classroom assignments and student achievement* (CSE Rep. 582). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

MATSUMURA, L. C. & PASCAL, J. (2003). *Teachers' assignments and student work: Opening a window on classroom practice* (CSE Rep. 602). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

MATSUMURA, L. C., SLATER, S. C., WOLF, M. K., CROSSON, A., LEVISON, A., PETERSON, M., RESNICK, L. & JUNKER, B. W. (2006). *Using the instructional quality assessment toolkit to investigate the quality of reading comprehension assignments and student work* (CSE Rep. 669). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

MESSICK, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (13-103). New York: MacMillan.

NATIONAL COUNCIL ON EDUCATION STANDARDS AND TESTING. (1992). *Raising standards for American education*. Washington, DC: U.S. Government Printing Office. (ERIC Document Reproduction Service No. ED338721)

NIEMI, D., WANG, J., STEINBERG, D. H., BAKER, E. L. & WANG, H. (2007). Instructional sensitivity of a complex language arts performance assessment. *Educational Assessment*, 12(3&4), 215-237.

NO CHILD LEFT BEHIND ACT OF 2001, Pub. L. No. 107-110, § 115 Stat. 1425 (2002).

PELLEGRINO, J. P., CHUDOWSKY, N. & GLASER, R. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

POPHAM, W. J. & BAKER, E. L. (1970). *Systematic instruction*. Englewood Cliffs, NJ: Prentice-Hall.

POPHAM, W. J. & HUSEK, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 6, 1-9.

ROGOSA, D. (1999a). *Accuracy of individual scores expressed in percentile ranks: Classical test theory calculations* (CSE Tech. Rep. No. 509). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

— (1999b). *Accuracy of Year-1, Year-2 comparisons using individual percentile rank scores: Classical test theory calculations* (CSE Tech. Rep. No. 510). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

— (1999c). *How accurate are the STAR national percentile rank scores for individual students? An interpretive guide*. Palo Alto, CA: Stanford University.

STECHER, B., BORKO, H., KUFFNER, K. L., MARTINEZ, F., ARNOLD, S. C., BARNES, D., CREIGHTON, L. & GILBERT, M. L. (2007). *Using artifacts to describe instruction: Lessons learned from studying reform-oriented instruction in middle school mathematics and science* (CSE Rep. 705). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

SWELLER, J. (1999). *Instructional design in technical areas*. Camberwell, Australia: ACER Press.

TOBIAS, S. (1976). Achievement treatment interactions. *Review of Educational Research*, 46, 61-74.

VENDLINSKI, T. P., BAKER, E. L., & NIEMI, D. (2008). Templates and objects in authoring problem-solving assessments. In E. Baker, J. Dickieson, W. Wulfeck, & H. F. O'Neil, (Eds.), *Assessment of problem solving using simulations* (309-333). New York: Erlbaum.

VENDLINSKI, T. & STEVENS, R. (2000). The use of artificial neural nets (ANN) to help evaluate student problem solving strategies. In B. Fishman & S. O'Conner-Divelbiss (Eds.), *Proceedings of the fourth international conference of the learning sciences* (108-114.). Mahwah, NJ: Erlbaum.

WEBB, N. L. (1999). *Research Monograph No. 18: Alignment of science and mathematics standards and assessments in four states*. Madison, WI: National Institute for Science Education.

YOON, B. & RESNICK, L. (1998). *Instructional validity, opportunity to learn and equity: New standards examinations for the California mathematics renaissance* (CSE Rep. 484). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Electronic Resources

BAKER, E. L. (2007a, August/September). The end(s) of testing (2007 AERA Presidential Address). Educational Researcher, 36(6), 309-317. Retrieved October 2, 2007, from http://www.aera.net/uploadedFiles/Publications/Journals/Educational_Researcher/3606/09edr07_309-317.pdf

OECD. (2005). *PISA 2003 data analysis manual: SAS® users*. Paris: Author, from http://www.pisa.oecd.org/dataoecd/53/22/35014883.pdf

Contact: Eva L. Baker. Universidad de California, Los Angeles (UCLA). National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Los Angeles California, USA. E-mail: eva@ucla.edu