Using Instructional Sensitivity and Instructional Opportunities to Interpret Students' Mathematics Performance

Marsha Ing
Stanford University

Abstract
Within the context of a mathematics reform effort to implement algebra in elementary schools, there is pressure to provide results in the form of student achievement scores. However, widely-used measures of student achievement may be unrelated to the ideas and instructional practices encouraged by the reform effort. The inappropriate use of student achievement scores often leads to inaccurate inferences about the quality of instruction. This study explores the validity of inferences about instructional quality using two measures of mathematics achievement: a measure of algebraic reasoning designed to closely relate to instructional activities and a measure of grade-level specific California content standards. This exploration includes multiple measures of classroom instruction to evaluate the instructional sensitivity of multiple measures of math achievement and applies an analytic method that makes it possible to relate student-level outcomes to teacher-level measures of instruction. Findings suggest that particular items measuring equality and relational thinking from the measure of algebraic reasoning were sensitive to instruction. The ability of these measures to determine the impact of instruction on student performance depends on the variables that define students' opportunities to learn and the characteristics of the student assessment items. These factors should be considered when evaluating the relationship between instructional quality and student performance. Instructional sensitivity provides a framework to interpret student performance by creating a link between instructional opportunities and performance on particular assessment items.

Introduction

Student performance on achievement tests is often used to make inferences about the content and quality of instruction students have received. If test scores are high or show improvement over time, instructional quality is assumed to be good or improving. Conversely, if test scores are low or decrease over time, instructional quality is assumed to be poor or declining. This assumption is at the heart of conclusions made about educational quality from test scores and is captured in the following statement by the executive director of Education Trust West in a *Los Angeles Times* article (Vaughn, 2005) concerning California's scores on the National Assessment of Educational Progress: "No matter how you look at this data, California is at the bottom. There is something systematically wrong with the way we are approaching educating all students in this state" (p. A20).

The assumption that something is wrong with education is based on sparse information about classroom instruction or little knowledge of the implementation of various reform efforts. This leads to imprecise evaluations of the quality of education and the effectiveness of educational reform. Such inferences about educational reform assume that test scores adequately reflect instructional content and quality (Airasian & Madaus, 1983; Airasian & Miranda, 2002; Amrein & Berliner, 2002; Moss, Pullin, Gee, & Haertel, 2005), that is, that test scores are sensitive to instruction. The term instructional sensitivity (Burstein, 1983, 1989; Haladyna & Roid, 1981; Miller & Linn, 1988; Popham, 2006; Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002), then, describes a property of an assessment that addresses these assumptions and inferences.

A Study of Instructional Sensitivity

The purpose of this study is to use the concept of instructional sensitivity to explore the validity of inferences about instructional quality, and to explain student performance on two different assessments. One of the assessments was designed to measure algebraic reasoning (Carpenter, Franke, & Levi, 2003; Carpenter & Levi, 2004; Carpenter, Levi, Berman, & Pligge, 2005) and the other to measure California math content standards.

An item classified as instructionally sensitive describes the degree to which it "reflects student knowledge/ability as the consequence of instruction"

(Burstein, 1989, p. 5). Burstein (1983) emphasizes that:

> An exact explanation of how a student responds to given test items is unanswerable under all but the most trivial circumstances. Nonetheless, it is reasonable to attempt to narrow the range of plausible explanations and to investigate the likelihood that particular instructional experiences activate processes that account for student responses. (p. 99)

While we will never know exactly why students respond to an item in a particular way, instructional sensitivity is a quality of an assessment that can be used to help explain performance of students with different instructional experiences.

Among students with similar ability levels, if students with one set of instructional opportunities perform better on an item than students with a different set of instructional opportunities, the item is considered to be sensitive to the effects of instruction. If students with similar ability levels who have different instructional experiences all perform the same on an item, the item is considered to be insensitive to the effects of instruction. In the latter case, student performance on an item does not depend on instructional experiences but depends instead on other factors such as student ability level or earlier achievement.

Educators recognize that not all assessments are designed to have the same degree of instructional sensitivity. Some assessments are meant to be measures of general achievement not necessarily influenced by particular instructional opportunities while other assessments are designed as measures of student understanding of a specific lesson and are assumed to be influenced by the specific types of instructional opportunities provided. Furthermore, the instructional opportunities that predict performance on the assessment of a particular lesson (close proximity to instruction) might not predict performance on the assessment of general achievement (more distant from instruction).

In considering the instructional sensitivity of different kinds of assessments, it is useful to consider the multilevel framework for evaluating the instructional sensitivity of different assessments proposed by Ruiz-Primo et al. (2002). Their framework (Figure 1) considers student performance on assessments that vary in terms of proximity from instruction provided.

It is possible is to draw conclusions about student proficiency based on assessments that are closer to instruction. However, researchers have cautioned about focusing only on measures close to instruction (Cronbach, 1963; Koretz,

1996; Linn, 1983; Mehrens & Ebel, 1979; Ruiz-Primo et al., 2002; Shepard, 2004). Cronbach (1963) suggested that "an ideal evaluation would include measures of all the types of proficiency that might reasonably be desired in the area of question, not just the selected outcomes to which this curriculum directs substantial attention" (p. 680). Linn (1983) continued the argument that "such a tight link of the test to what is taught has considerable appeal" and is "apt to yield results that are more sensitive to instruction that is taking place" (p. 186) but concluded that doing so is at the "risk of reduction in the importance of learning to apply skill and knowledge to new problems" (p. 187). Thus, central to investigating the effects of educational reform is the inclusion of assessments at different proximities to instruction.

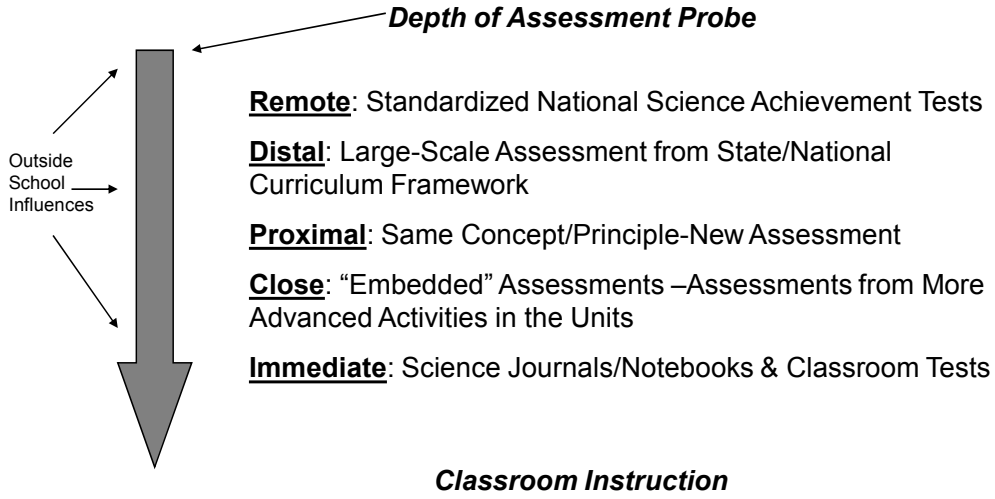

Figure 1. Characterization of Assessments Based on Proximity to Classroom Instruction from Ruiz-Primo et al. (2002).

Literature Review

Instructional Sensitivity and Related Concepts

The increased use of test scores as measures of educational quality calls for greater attention to the types of measures used to make such judgments (National Research Council, 2001a, 2001b). Less attention has focused on how these different assessments can be used to create more comprehensive judgments about educational quality. Instructional sensitivity helps focus attention on the validity of these different measures by providing guidance about which items are sensitive to the effects of instruction. Instructional

sensitivity is related to concepts such as alignment (e.g., Porter, 2002; Rothman, 2004, Webb, 1997, 2002;), opportunity to learn (e.g., Anderson, 1985; Floden, 2002; Guiton & Oakes, 1995; Herman, Klein, & Abedi, 2000; Wang, 1998; Wiley & Yoon, 1995), and test preparation (e.g., Anastasi, 1981; DerSimonian & Laird, 1983; Herman & Ing, 2007; Messick, 1982) all of which seek to strengthen the validity of inferences about instructional quality based on test performance.

Compared to opportunity to learn (OTL) and alignment, instructional sensitivity, however, is more narrowly defined and focused. As an example of this narrow definition, the term instructional validity is used by Yoon and Resnick (1998) to describe a quality of the assessment that "is systematically sensitive to differences in opportunity to learn" and "registers differences in the amount and kind of instruction to which students have been exposed" (p. 2).

Different Approaches to Measuring Instructional Sensitivity

Early approaches to measuring instructional sensitivity did not include information about instruction. A typical approach flagged response patterns that seemed unusual or unexpected (Donlon & Fischer, 1968; Hanna & Bennett, 1984; Harnisch, 1983; Kane & Brennan, 1980; Sato, 1975; Tatsuoka & Tatsuoka, 1980). An unusual response pattern was identified for profiles of responses with the same total score but different responses to particular items. For example, a large value of Sato's caution index (1975) indicated an unusual response pattern and served as a caution against the use of the total score as an accurate measure for a particular examinee. An examinee who answered 8 out of 10 items correctly was expected to answer all of the easy items correctly but miss the 2 most difficult items. However, if the examinee with a score of 8 out of 10 missed the 2 easy items but answered all of the other items correctly, Sato's caution index would be high. Studies used these unusual patterns to flag schools or students who might be randomly responding to items. Researchers suggested that random responses could be due to test anxiety or carelessness but might also indicated that respondents were not instructed on the test material. If respondents were not instructed in the material, their responses might be more random than those of instructed students and might not follow the usual pattern of responses.

Harnisch and Linn (1981) compared several of these unusual response pattern indices for seven different schools and hypothesized that these unusual

response patterns might be a result of school variability in content coverage and emphasis as well as attendance patterns. The researchers did not, however, obtain measures of content coverage and emphasis nor any other indicators of school variability pertaining to instructional quality or opportunities to learn.

This early approach to studying instructional sensitivity finds a parallel in a recent measurement approach to explaining differences in performance known as differential item functioning (Holland & Wainer, 1993) based on item response theory (e.g., Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Lord, 1980; Van der Linden & Hambleton, 1997). Item response theory is a psychometric approach to estimating an unobservable trait based on item responses. Differential item functioning occurs when respondents with similar trait levels have a different probability of responding correctly to a given item. The item that demonstrates differential item functioning is thought to be biased toward a particular group of students. In terms of instructional sensitivity, if students with the same trait level are given different instructional opportunities, an item would be biased toward students who were provided with instruction.

An example of this can be found in a study by Clauser, Nungester, and Swaminathan (1996) who used differential item functioning to detect unusual response patterns due to educational experiences. Their initial analysis of data from the National Board of Medical Examiners indicated differences between males and females with similar ability levels on over 130 items. The authors hypothesized that this gender bias might be explained by differential educational opportunities of the respondents. When the authors included information about the residency fields (internal medicine, surgery, pediatrics, or obstetrics-gynecology) in their analyses, they found a decrease in the number of items displaying differential item functioning. The authors suggested that gender is confounded with educational experiences because gender and choice of residency fields are related variables. Thus, attributing differences in item performance to gender alone is misleading. Although Clauser et al. (1996) did not measure instructional opportunities directly, their study suggests that instructional information can be used to identify unusual response patterns.

Other early approaches to measuring instructional sensitivity used a pre-post test design to compare instructed and non-instructed students (Brennan & Stolurow, 1971; Haladyna & Roid, 1981; Popham, 1971). As described in

Haladyna and Roid (1981), Cox and Vargas (1972) calculated the difference between the difficulty level of the item at the time of the pretest and the posttest when given to students who were instructed and students who were not instructed. The difficulty level of the items on the post-test was expected to be lower for students who were provided instruction than students who were not provided with any instruction. That is, items should appear easier for students who have experience with the material. The authors found differences in the item difficulty levels between instructed and non-instructed students and attributed these differences to instructional sensitivity of the items. While these early approaches seemed to confirm expectations about change in performance between groups of students with different instructional experiences, they did not include information about the actual instructional experiences.

Later studies of instructional sensitivity included more explicit measures of instruction, such as content coverage, measured using student or teacher questionnaires about the opportunities provided to students (Miller & Linn, 1988; Muthén, 1989a, 1989b, 1994; Muthén, Huang, Khoo, Goff, Novak, & Shih, 1995; Muthén, Kao, & Burstein, 1991), textbook analysis (Mehrens & Phillips, 1986, 1987; Phillips & Mehrens, 1987) or assigned students to particular instructional conditions (Hanson, McMorris, & Bailey, 1986).

For example, Muthén and colleagues (Muthén, 1989a, 1989b, 1994; Muthén et al., 1995; Muthén et al., 1991) used more explicit measures of instruction from the Second International Mathematics Study (SIMS) in their studies on instructional sensitivity. In a 1994 article, Muthén describes the instructional measures used in his modeling of eighth-grade student performance on the SIMS mathematics assessment. Teachers were asked two questions regarding each of the items on the SIMS mathematics assessment: (a) During this school year did you teach or review the mathematics needed to answer the item correctly; and (b) If in the school year you did not teach or review the mathematics needed to answer this item correctly, was it mainly because it had been taught prior to this school year; it will be taught later; it is not in the school curriculum at all; or for other reasons? Muthén (1994) found that these instructional questions had a smaller effect in predicting math achievement (as defined by performance on the SIMS math items) than did prior achievement and demographics such as gender and father's level of education. Given the framework presented by Ruiz-Primo et al. (2002) on measuring instructional sensitivity, the SIMS items would be considered

remote items. By design, *remote* items are far removed from particular instruction experiences and are not easily influenced by instruction. Muthén's approach to measuring instructional sensitivity, however, was unique in the way that it included attention to content coverage and used this information to predict performance.

Ruiz-Primo et al. (2002) studied instructional sensitivity by including multiple achievement measures with different proximities to instruction to evaluate science reform effort using student work as a measure of instruction, and administering pre-tests and post-tests for the close and proximal achievement measures. For some curriculum units, changes between pre- and post-test total scores (not item level) indicated that the close assessments were more sensitive to instruction than were the proximal assessments. Ruiz-Primo et al.'s approach of using a direct measure of instruction is an important feature. Science notebooks were collected from students in each classroom as a sample of student work. Students used their science notebooks as a record of class activities which were then used by researchers as an indicator of content coverage. This direct measure of instruction was assumed to be reflective of the types of activities that actually occurred in the classrooms.

Considering whether experiences in particular classrooms influenced student performance could be explored by looking at performance for each classroom and taking into consideration the shared experiences of students in that particular classroom. Researchers have called for attention to the hierarchical structure of educational data (e.g., Burstein, 1980; De Boeck & Wilson, 2004; Raudenbush, 1988; Raudenbush & Bryk, 2002; Seltzer, 2004). Previous research on instructional sensitivity acknowledged the nested nature of the data but did not incorporate this concern into the analyses. For example, as described earlier, Harnisch and Linn (1981) used grouping variables to describe differences in performance between different schools or different regions. Effects of group membership were estimated separately, rather than simultaneously as can be carried out through multilevel analysis. The analysis of variance as carried out by Harnisch and Linn compared overall group differences and did not provide information about what contributes to group differences, nor did it consider shared experiences of students within these groups that might explain group differences. Through a simultaneous multilevel analysis, the unique effects on performance as a result of students within classrooms or classrooms within districts are estimated. In other words, questions about particular teacher level characteristics could be used to address

differences in student performance.

Previous studies investigating the instructional sensitivity of assessments suggest a link between the proximity of the assessment to instruction and the sensitivity of the assessment to instruction. To further clarify this link depends on (a) having in-depth and detailed measures of instruction; (b) administering multiple measures of student performance; and (c) carrying out analyses that directly investigate the link between instruction and student performance on the assessments. The present study addressed these three issues by using a multilevel approach to account for performance of students nested within classrooms. At the first level, referred to as Level 1 or the student-level, information about student general achievement was used to predict performance on the proximal and distal measures. At the second level, referred to as Level 2 or the teacher-level, information about classroom instructional opportunities was used to predict student performance.

Methods and Procedures

The data for this study were collected as part of a professional development effort in a large urban school district (Jacobs, Franke, Carpenter, Levi, & Battey, 2007). The current study focused on 24 third grade teachers who participated in a year-long mathematics professional development program on algebraic reasoning. Algebraic reasoning was defined as "generalizing and formalizing patterns and regularities" (p. 259). Researchers involved in the professional development activities further conceptualized algebraic reasoning as "woven throughout the curriculum so that teachers viewed algebraic reasoning as pervading the mathematics curriculum rather than simply as one more topic to teach" (p. 260).

Measures

The two different student-level outcome measures (proximal and distal items) were collected at the end of the school year. In addition, general measures of student prior achievement collected at the end of the previous school year were used as covariates in the multilevel analyses. The proximal items were assumed to be close to instruction in that they matched the types of things that occurred in the classroom in terms of content emphasis and cognitive demand. In this study, proximal items were designed to closely relate to the ideas and

processes advocated by the professional development program. Items from a district-wide assessment were considered distal items. These grade-specific assessments were administered to all students in the district and targeted particular state math content standards. These items were assumed to be less related to particular instructional opportunities provided but similar to the proximal items in terms of content. The disattenuated correlations between the proximal items, distal items, and general measures of student achievement suggested that these measures were moderately related. Multilevel regression analyses were carried out separately for each of the two outcome measures.

The instructional opportunity variables collected from each teacher and the average classroom prior achievement were the classroom level variables included in this study to predict performance on the outcome measures. For this study, the following kinds of information about instructional opportunities were collected from a written teacher assessment and an oral teacher interview: teacher perceptions of students' opportunities to learn algebraic reasoning, teacher algebraic reasoning content knowledge, teacher confidence in their own algebraic reasoning content knowledge, and teacher knowledge of their students' algebraic reasoning strategies. This information represented different but related dimensions of teacher practice that previous research suggested as important to capture. These measures of instructional opportunities served as proxies of the instruction opportunities that students actually experienced in each classroom. There was no evidence to validate the extent to which these measures captured differences in the instructional opportunities in each classroom but the risk of this assumption was taken to explore instructional sensitivity.

Sample

The initial sample for this study included 486 students across 24 third-grade classrooms (24 teachers), and eight schools. These eight schools were fairly similar in terms of their academic performance, percent of students receiving free or reduced lunch and percent of students designated as English language learners. All of these schools served predominantly African American and Hispanic student populations.

Results

The correlation between prior student achievement and the outcome measures (Table 1) was highest for the distal items, $r = .61$, $p < .01$. The correlations between student prior achievement and the equality items from the proximal measure were moderate, $r = .30$, $p < .01$, but lower compared to the correlation between prior achievement and performance on the distal items, $r = .44$, $p < .01$. This suggests that the relationship between prior achievement and performance on the proximal items was less than the relationship between prior achievement and performance on the distal items.

Table 1
Correlations Between Prior Achievement, Instructional Opportunity Variables and Performance on Proximal and Distal Items

Variable	Proximal	Distal
Student-level ($n = 321$)		
Prior student achievement	0.30*	0.61*
Classroom-level ($n = 24$)		
Opportunity to learn	0.28	0.05
Content knowledge	0.29	0.24
Confidence	0.61*	0.23
Awareness of student strategies	0.59*	0.30
Prior average classroom achievement	-0.06	0.57*

Note. *$p < .01$.

The correlation between teacher awareness of students' strategies and performance on the proximal items was positive and moderate, $r = .59$, $p < .01$. The correlation was low and not significant between teacher awareness of students' strategies and the distal items, $r = .30$, $p > .01$. None of the other instructional opportunity variables correlated with performance on the proximal or distal items.

Predicting Performance on the Proximal Items

The final conditional multiple level linear regression model included prior achievement at the student-level and teacher awareness of students'

relational thinking strategies at the classroom-level. Average classroom prior achievement and other classroom-level instructional opportunity variables were not included in the final conditional model because these variables were not significantly correlated with performance on the equality items and were not significant predictors of performance in the multilevel models. Teacher confidence correlated with performance on the equality items but was excluded because it was highly correlated with teacher awareness of student strategies, $r = .78$, $p < .01$. Teacher awareness of student strategies was a focal point in the professional development program so this variable was selected for substantive reasons.

The fixed effect results for the final model (Table 2) indicated that the effect of student level prior achievement and teacher awareness of student strategies on performance on the equality items were small (less than one point) but significant. Teacher awareness of student strategies was not a significant predictor of the within-class achievement slopes. In other words, student prior achievement was an important consideration in student performance on the equality items. Teacher awareness of student strategies was also predictive of the average class means on the equality items but not on within-class differences.

Table 2
Fixed Effects for the Final Conditional Model of Student Performance on the Equality Items

Fixed Effect	Coefficient	*SE*	*t* ratio
Model for classroom means			
Intercept (γ_{00})	2.20	0.17	- - -
Awareness of student strategies (γ_{01})	0.58	0.11	5.17*
Model for achievement slopes			
Intercept (γ_{10})	0.71	0.11	- - -
Awareness of student strategies (γ_{11})	-0.01	0.07	-0.15

Note. *$p < .01$.

A smaller percentage of the variance is attributed to classroom differences compared to the unconditional model. This percentage shifted from 26% in the unconditional model to 24% in the final conditional model. The final results

for grade 3 indicated that there were two variables that helped explain average class performance on the equality items: student level prior achievement and teacher awareness of students' relational thinking strategies. A majority of the variation in performance on these items was due to differences between students. A smaller, but still significant, proportion of the variation was due to differences between classrooms.

Predicting Performance on the Distal Items

The multilevel models predicting performance on the distal items did not include any of the classroom-level instructional opportunity variables because none of them significantly correlated with performance. Student prior achievement and average classroom achievement significantly predicted performance but none of the instructional opportunity variables were predictive of performance on the distal items. The variance between students and variance between classrooms did not indicate that this student outcome data was very different from other achievement outcomes (Raudenbush, Martinez, & Spybrook, 2007) but these results suggest that the same instructional opportunity variables were not equally predictive of performance on the different student outcomes.

Discussion and Conclusions

Mislevy, Wilson, Ercikan, and Chudowsky (2003) describe educational assessments as "data that becomes evidence in some analytic problem only when we have established their relevance to some conjecture we are considering" (p. 495). Educational assessments are often taken as evidence of instructional quality without attention to whether and how instruction and student performance are related. This study attempted to address this issue by exploring the link between instructional opportunities and performance on multiple measures of math achievement using the concept of instructional sensitivity. As Popham (2006) proposed, "instructional sensitivity is best conceived of as a continuum rather than a dichotomy" and that "rarely will one encounter an accountability test that is totally sensitive or totally insensitive to instruction" (p. 3).

Limitations

This study addressed this notion that a particular test is better (or worse) than another for making judgments about instructional quality by exploring the power of subscores from each test to distinguish between different instructional opportunities. There was limited evidence to suggest that performance on items on a proximal measure was influenced by instructional opportunities compared to items on a distal measure. Evidence was limited to third graders participating in this study and instruction opportunities are narrowly defined as teacher awareness of students' relational thinking strategies. There are several possible reasons why the instructional opportunity variables, other than teacher awareness of relational thinking as a possible strategy students might use to solve problems, failed to significantly predict student performance on the proximal and distal items.

First, instructional opportunities simply do not matter in terms of student performance on these items. While this is certainly a possibility, one might argue that a premise or goal of instruction is to influence student performance over and above their prior achievement. Second, in addition to instructional opportunity variables that related to equality and relational thinking, opportunities to learn the items on the distal measure were not available but should be included in future analyses. A final possible reason for the lack of relationships is that there were restricted ranges in the responses to the instructional opportunity variables. For example, all teachers included in this study identified relational thinking strategies their students might use to solve problems. This select sample of teachers might not represent the entire population of teachers in terms of these variables and thus would only represent the relationship between instructional opportunities and performance for a limited portion of the population. Broadening the sample of teachers to create a greater range of responses (beyond the teachers participating in the professional development program), might address the limited range of talent available for this study.

Future Directions

This study raised broader issues that push researchers to think more richly about how instructional opportunities are defined. The assumption that previous research has led us to believe is that given the best possible measures of instruction and the appropriate analytic methods, empirical evidence for the

relationship between instruction and student performance would be clear. By necessity, quantitative measures of instruction will always be approximations of the actual instruction that occurs in each classroom (e.g., Burstein, McDonnell, Van Winkle, Ormseth, Mirocha, & Guitton, 1995). In other words, it is obvious that these large-scale measures will always be indirect measures of instruction and will be reductive in some way. Thus, the goal of creating these measures is to capture key features of instructional opportunities *and* to articulate the limitations of these particular measures. In a large scale study of the relationship between instruction and student performance, it is unlikely and logistically overwhelming to observe key features of instruction in every classroom. Even if researchers observed instructional opportunities in every single classroom participating in the large scale study, it is difficult to make sense of all of this information in an efficient manner. Thus, the need to create quantitative measures of instruction is absolutely essential when conducting large scale quantitative studies. The goal when creating these measures is to come as close as possible to describing the key features of instructional opportunities and to acknowledge the limitations of these approximations.

Acknowledgements

The author would like to thank Noreen Webb of the University of California, Los Angeles, for her comments on an earlier version of this article.

References

Airasian, P. W., & Madaus, G. F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement, 20*(2), 103-118.

Airasian, P. W., & Miranda, H. (2002). The role of assessments in the revised taxonomy. *Theory into Practice, 41*(4), 249-254.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning *Education Policy Analysis Achieves, 10*(18). Retrieved October 1, 2007 from http://epaa.asu.edu/epaa/v10n18/

Anastasi, A. (1981). Coaching, test sophistication and developed abilities. *American Psychologist, 36*(10), 1086-1093.

Anderson, L. W. (1985). Opportunity to learn. In T. Husen & T.N. Postelethwaite (Eds.), *The international encyclopedia of education research and studies,* (*Volume 6*, pp. 3682-3686). New York: Pergamon Press.

Brennan, R. L., & Stolurow, L. M. (1971). *An empirical decision process for formative evaluation*. Research Memorandum No. 4. Cambridge: MA: Harvard University CAI Laboratory.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education, 8*, 158-233.

Burstein, L. (1983). A word about this issue [Editor's note]. *Journal of Educational Measurement, 20*, 99-102.

Burstein, L. (1989). *Conceptual considerations in instructionally sensitive assessment*. (CSE Technical Report 333). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Burstein, L., McDonnell, L. M., Van Winkle, J., Ormseth, T., Mirocha, J., & Guitton, G. (1995). *Validating national curriculum indicators*. Santa Monica, CA: RAND.

Carpenter, T. P., Franke, M. L., & Levi, L. (2003). *Thinking mathematically: Integrating arithmetic and algebra in elementary school*. Portsmouth, NH: Heinemann.

Carpenter, T. P., & Levi, L. (2004). *Developing conceptions of algebraic reasoning in the primary grades*. (National Center for Improving Student Learning and Achievement in Mathematics and Science Research Report No. 00-2). Madison, WI: NCISLA, Wisconsin Center for Education Research.

Carpenter, T. P., Levi, L., Berman, P. W., & Pligge, M. (2005). Developing algebraic reasoning in elementary school. In T. A. Romberg, T. P. Carpenter, & F. Dremock (Eds.), *Understanding mathematics and science matters* (pp. 81-98). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background. *Journal of Educational Measurement, 33*(4), 453-464.

Cox, R. C., & Vargas, J. (1972). *A comparison of item selection techniques for norm-referenced and criterion-referenced tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Cronbach, L. J. (1963). Course improvement through evaluation. *Teachers College Record, 64*(8), 672-683.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.

DerSimonian, R., & Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harvard Educational Review, 53*, 1-15.

Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group determined item difficulties. *Educational and Psychological Measurement, 28*, 105-113.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Floden, R. E. (2002). The measurement of opportunity to learn. In A.C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 231-267). Washington, DC: National Research Council.

Guiton, G., & Oakes J. (1995). Opportunity to learn and conceptions of educational equality. *Educational Evaluation and Policy Analysis, 17*(3), 323-336.

Haladyna, T. M., & Roid, G. (1981). The role of instructional sensitivity in the empirical review of criterion-referenced test items. *Journal of Educational Measurement, 18*(1), 39-53.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijoff Publishers.

Hanna, G. S., & Bennett, J. A. (1984). Instructional sensitivity expanded. *Educational and Psychological Measurement, 44*, 583-596.

Hanson, R. A., McMorris, R. F., & Bailey, J. D. (1986). Differences in instructional sensitivity between item formats and between achievement test items. *Journal of Educational Measurement, 23*(1), 1-12.

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. *Journal of Educational Measurement, 20*(2), 191-206.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18*(3), 133-146.

Herman, J., & Ing, M. (2007). Test Preparation. In K. M. Borman, S. E. Cahill, & B. A. Cotner (Eds.), *The Praeger handbook of American high schools* (pp. 416-420). Westport, CT: Praeger.

Herman, J. L., Klein, D. C., & Abedi, J. (2000). Assessment students' opportunity to learn: Teacher and student perspectives. *Educational Measurement: Issues and Practice, 19*(4), 16-24.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning: Theory and practice*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jacobs, V. R., Franke, M. L., Carpenter, T. P., Levi, L., & Battey, D. (2007). Professional development focused on children's algebraic reasoning in elementary schools. *Journal for Research in Mathematics Education, 38*(3), 258-288.

Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement, 4*, 105-126.

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 Academy Press.

Linn, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement, 20*(2), 179-189.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Erlbaum.

Mehrens, W. A., & Ebel, R. L. (1979). Some comments on criterion-referenced and norm-referenced achievement tests. *NCME, Measurement in Education, 10*(1), 1-8.

Mehrens, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. *Journal of Educational Measurement, 23*(3), 185-196.

Mehrens, W. A., & Phillips, S. E. (1987). Sensitivity of item difficulties to curricular validity. *Journal of Educational Measurement, 24*(4), 357-370.

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist, 17*(2), 67-91.

Miller, M. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement, 25*(3), 205-219.

Mislevy, R. J. Wilson, M., Ercikan, K., & Chudowsky, N. (2003). Psychometric principles in Student Assessment. In, T. Kellaghan, & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation*, (pp. 489-532). Dordrecht, The Netherlands: Kluwer Academic Press.

Moss, P. A., Pullin, D., Gee, J. P., & Haertel, E. H. (2005). The idea of testing: Psychometric and sociocultural perspectives. *Measurement, 3*(2), 63-83.

Muthén, B. O. (1989a). Using item-specific instructional information in achievement modeling. *Psychometrika, 54*(3), 385-396.

Muthén, B. O. (1989b). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*(4), 557-586.

Muthén, B. O. (1994). Instructionally sensitive psychometrics: Applications to the Second International Mathematics Study. In I. Westbury, C. A. Ethington, L. A. Sosniak, & D. P. Baker (Eds.), *In search of more effective mathematics instruction* (pp. 293-324). Norwood, NJ: Ablex Publishing Corporation.

Muthén, B. O., Huang, L. C., Khoo, S. K., Goff, G. H., Novak, J. R., & Shin, J. C. (1995). Opportunity-to-learn effects on achievement: Analytical aspects. *Educational Evaluation and Policy Analysis, 17*(3), 371-403.

Muthén, B. O., Kao, C. F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*, 1-22.

National Research Council. (2001a). *Classroom assessment and the National Science Education Standards*. Committee on Classroom Assessment and the National Science Education Standards. J. M. Atkin, P. Black, & J. Coffey (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (2001b). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundation of Assessment. J. Pellegrino, N. Chudowsky and R. Glaser (Eds.). Washington, DC: National Academy Press.

Phillip, S. E., & Mehrens, W. A. (1987). Curricular differences and unidimensionality of achievement test data: An exploratory analysis. *Journal of Educational Measurement, 24*(1), 1-16.

Popham, W. J. (1971). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Educational Technology Publications.

Popham, W. J. (2006). *Determining the instructional sensitivity of accountability tests*. Paper presented at the annual Large-Scale Assessment Conference, Council of Chief State School Officers, San Francisco, CA.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3-14.

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics, 13*(2), 85-116.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29*, 5-29.

Rothman, R. (2004). Benchmarking and alignment of state standards and assessments. In S. H. Furman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 96-114). New York: Teachers College Press.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L. S., & Klein, S. (2002). On the evaluation of systematic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39*(5), 369-393.

Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo, Japan: Meiji Tosho.

Seltzer, M. H. (2004). The use of hierarchical models in analyzing data from experiments and quasi-experiments conducted in field settings. In D. Kaplan (Ed.), *The Handbook of Statistical Methods for the Social Sciences* (pp. 259-280). Thousand Oaks, CA: Sage Publications.

Shepard, L. (2004). Curricular coherence in assessment design. In M. Wilson (Ed.), *Toward coherence between classroom assessment and accountability* (pp. 239-249). Chicago: National Society for the Study of Education.

Tatsuoka, K., & Tatsuoka, M. M. (1980). *Detection of aberrant response patterns and their effects on dimensionality* (Research Report 80-4.). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York: Springer.

Vaughn, E. (2005, October 20). California students are still struggling: Reading and math test scores for fourth- and eighth-graders rank near the bottom in the nation. *Los Angeles Times*, pp. A20.

Wang, J. (1998). Opportunity to learn: The impacts and policy implications. *Educational Evaluation and Policy Analysis, 20*(3), 137-156.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Madison, WI: National Institute for Science Education and Council of Chief State School Officers.

Webb, N. L. (2002). *An analysis of the alignment between mathematics standards and assessments for three states.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Wiley, D. E., & Yoon, B. (1995). Teacher reports on opportunity to learn: Analyses of the 1993 California Learning Assessment System (CLAS). *Educational Evaluation and Policy Analysis, 17*(3), 355-370.

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