Conceptual Considerations in Instructionally Sensitive Assessment

Leigh Burstein

CSE TECHNICAL REPORT 333

Center for Research on Evaluation, Standards, and Student Testing
University of California, Los Angeles

Presented at AERA Annual Meeting, San Francisco, March 31,1989

The presentations that constitute this symposium describe work on the Instructional Assessment Project at the Center for Research on Evaluation, Standards, and Student Testing (CRESST) under a grant from the Office of Educational Research and Improvement (OERI). The overall goals of the project are (1) to improve the validity of subject matter tests and (2) to improve the instructional sensitivity of psychometric and statistical methods used to analyze, interpret, and report test data intended to guide instructional and curriculum decision-making. The presentations focus on work intended to contribute to improvements in the conceptual and technical foundations of instructionally sensitive assessment within the context of large-scale achievement testing.

The conceptual and analytical studies reported here draw upon secondary analyses of U.S. data from the Second International Mathematics Study (SIMS; Burstein et al., in progress; McKnight et al., 1987; Robitaille and Garden, 1989; Travers & Westbury, 1989). In 1981-82, 180 mathematics achievement items were administered on two occasions (near the beginning and end of the school year) to students in over 200 8th grade mathematics classes (remedial, typical, enriched, algebra) which varied considerably in opportunity to learn the topics covered in the test, use of textbooks, and use of a variety of teaching styles. The SIMS data set is unusually rich in its set of teacher-reported information about classroom activities, including item-specific opportunity to learn information. Overall, this data set incorporates the kinds and range of information deemed necessary to develop and investigate models of student knowledge/ability as a consequence of their instructional experiences.

My presentation describes the conceptual framework for instructionally sensitive assessment that guided the analyses of data from the Second International Mathematics Study (SIMS). Both the intellectual rationale for the project effort and the current draft version of the guiding conception of "instructionally sensitive psychometrics" are discussed. I will also consider briefly certain practical realities derived from our experiences in analyzing SIMS thus far and possible implications for applications of the methodology in other large-scale achievement contexts.

Instructional Assessment

Testing has played a major role in the many educational reforms that have been enacted throughout the country in the past few years (Linn, 1987; Pipho, 1985). Clearly, policy makers expect that new test requirements will provide a mechanism for improving instruction and student learning. As Petrie (1987) has put it, tests are the "engines of educational reform" (p. l).

The contrast between these expectations and the views of many classroom teachers and educational researchers about the value of tests for improving instructional decisions is stark. Many teachers find that stadardized tests provide them with little diagnostic information that can be used to inprove their instructional decisions. There is also a paucity of research evidence supporting the claim that tests provide valid diagnostic information or the notion that test results have improved the day-to-day decison making in the classroom (Baker, 1987; Baker & Herman, 1986; Bejar, 1984; Burstein, 1983; Glaser, 1986; Linn, 1986).

There are, of course, many reasons for the contrasting perspectives about the value of tests in improving instruction and student learning. To begin with, the perspectives, while quite different, are not necessarily contradictory. Tests that have been introduced in most of the reform legislation need not provide diagnostic information or influence day-to-day decision making in order to affect teaching and learning. When important consequences are perceived to be attached to test results, the tests can have a major influence over topic coverage and emphasis (see, for example, Airasian & Madaus, 1983; Linn, 1988). The change in emphasis can create a better match between what is taught and what is tested, and, as a consequence, affect test performance (Cooley & Leinhardt, 1980; Leinhardt, 1983; Leinhardt & Seewald, 1981).

Policy makers rightly assume that test requirements can lead to other educational changes that, in turn, produce changes in student performance. The required tests send a clear message to teachers about what is considered important, and thereby shape the curriculum. They can do this, however, without teachers using the test results of individual pupils to make instructional decisions. The potential of tests to serve the latter function is largely unrealized. How and, indeed whether, testing systems can be devised that contribute to instruction and student learning in this way remain important, but largely unanswered questions that are the focus of the Instructional Assessment project.

The work reported here is concerned with the improvement of the instructional sensitivity of psychometric and statistical methods used to analyze, interpret, and report test data intended to guide instructional decisions. Work already completed on the project suggests a number of important factors which influence the validity of the use of test results for instructional decisions and instructional sensitivity of the tests (Burstein, Webb, Muthen, & Linn, 1987).

First, it is evident that there are multiple, systematic factors that contribute to student performance as measured by an instructional assessment at a given point in time. In order for such measurements to effectively aid in instructional improvement, they must be designed, analyzed, and reported in ways that distinguish among the factors (student ability, topic exposure, forms of instructional exposure) that affect performance (Burstein, Chen, Sen & Kim, 1987; Muthen, Kao & Burstein, 1987; Shavelson, Webb, Shemesh & Yang, 1988). Otherwise the validity of the assessment for the purposes it is intended to serve is likely to be compromised or weakened. For example, it is a waste of resources to attempt individually-targeted remediation when the basis for a particular diagnostic signal is lack of exposure to the knowledge and skills assessed.

Second, auxiliary information about characteristics of the students and their instructional experiences are essential to properly utilizing appropriately designed instructional assessments. Student characteristic data can serve various purposes. Prior ability and performance data can help distinguish learning within a given setting and whether students from all levels benefit proportionately or otherwise. Gender, ethnicity, and other socio-demographic information allow the possibility of monitoring equity considerations. The effort to remedy the social and structural conditions that allow these factors to influence educational opportunities and, hence, performance requires a conscious attempt to monitor these factors in instructional assessment efforts.

With respect to instructional experiences, minimally, the ability to distinguish among the different educational settings (classes and schools) in which assessments are administered is a necessary condition for an appropriate interpretation of student performance data. Information about actual topic coverage and instructional methods are of even greater value.

Third, it is clear that the psychometric models and statistical analyses that have served traditional goals of measuring individual differences for purposes of prediction, are, at best, incomplete for purposes of identifying the instructional factors which influence performance on specific problems, identifying item content and process characteristics that are most sensitive to instructional practices, and the maximization of the instructional utility of achievement test results. Analytic procedures that incorporate information about student background, prior instructional experiences, item content, and the cognitive processes involved in answering questions (e.g., Muthen et al., 1988) are needed.

Fourth, beyond better analytic methods, there is a need for a better understanding of test items themselves. Among the pertinent questions in this are the influences of both alternative representations of problems and alternative forms of response on student cognitive processes and performance levels (Webb & Shavelson, 1987; Shavelson et al., 1988).

The issues highlighted are that appropriate models for student performance and proper contextualization through the collection of auxiliary information about students and their instructional programs are essential in the design, analysis and interpretation of multipurpose instructional assessments. Results from instructional assessments reflect both demographic and ability attributes of the students and the instructional intents, opportunities, and execution of the settings in which students reside; moreover, these accrue and change over time. Thus one needs to be cognizant of these circumstances and sensitive to their consequences for the validity of test interpretations.

One implication from the analytical techniques applied thus far is that it is important not to value complexity of analytical methods for their own sake in validating and applying instructional assessments. Sensitivity to the nature of the substantive processes that underly performance and the educational structure within which instruction takes place should be the primary criterion for judging the appropriateness and utility of analytical methods (Burstein, 1980, 1983; 1984, 1986; Shavelson, Webb, & Burstein, 1986; Sirotnik & Burstein, 1986). Contrary to what one might surmise from the prevailing literature, this is especially true when the focus of the assessment is to guide instructional improvement rather than to contribute to the educational research knowledge base.

Another point that comes through is the value of collecting assessment and accompanying characteristic data longitudinally. Benefits that accrued from the two measurement occasions in SIMS compared with the single cross-sectional data collection in other major national and international educational surveys are well documented (e.g., Burstein et al. , in press; Kifer, Schmidt & Wolfe, 1984, in press; Delandshere, 1986; Muthen et al., 1987, Wolfe, 1986)).

What is common to the analytical methods employed thus far in our own investigations is their use for analyzing patterns of performance. Whether one is dealing with a set of items designed to represent different topics or ways of understanding, or with students differentiated by their demographic attributes or the attributes of the instructional histories or with performance on a single measurement over time, it is important that the analysis methods mirror to the extent possible the structure of the educational situation. For example, if students in Algebra class A are not taught simple formulas at all, students in class B are taught using only symbolic manipulations of algebraic equations (e.g., if P=LW, and L=2 and P=6, what is W), and students in class C receive instruction that also embed the formula within real situations (e.g., word problems with and without accompanying pictorial representations), one would want a test designed to reflect the distinctions in the classes's experiences and an analysis that will detect these distinctions with sufficient specificity to point to areas of needed instructional improvement. And, presuming no prior introduction to the topic, one would want to employ analytical methods that associate performance patterns with class-level instructional differences rather than imply that student attributes are responsible for the differences.

Again, taking the studies as a set, certain characteristics of methodology that influence the validity of the design, analysis, and interpretation of instructional assessments are evident. First, analysis strategies that are sensitive to data structure are essential. These may be simple techniques but their use should follow a comprehensive conception of student performance in educational and social contexts within the course of their schooling. A corollary of the first point is that analyses of data only in the form of highly aggregated scores (total scores representing horizontal aggregations performance across topics and ways of understanding) and at only a single educational level (whether students only or at some level of vertical aggregation to classes, courses or course sequences, schools, districts, and so forth) are unlikely to be sufficiently informative to guide instructional improvement. Analyses, like designs, need to be multifaceted and multidimensional since attributes of any given performance certainly are.

"Instructionally Sensitive Psychometrics"

To gain some perspective on our approach to instructional sensitivity, it is useful to contrast our conception with one more prevalent in traditional psychometric practice. As a starting point, we argue that whether classical or modern (IRT), conventional psychometric analyses of educational achievement data, guided by a concern for valid measurement of individual differences in ability constructs, typically leave little room for capturing the complexity of educational experiences of students within naturally occurring educational settings.

Contrasting views about the relationship of the "experiences a student has" and "his/her performance/achievement/ability" is central to distinguishing traditional conceptions from our own. Traditionally, instructionally sensitive psychometric methods are those which lead to better student-level diagnoses. The focus is clearly on the "ability" attribute, a property of the individual student at some point in time. One is not concerned about how individual differences got to be the way they are but rather whether the assessment, and its accompanying analyses, lead to accurate and efficient diagnoses of individuals' levels of functioning on given achievement/ability constructs.

Our work has approached this problem from the opposite direction. That is, while we want to validly measure ability/achievement constructs and individual differences therein, our primary concern is on how ability/achievement measures reflect educational and other experiences, or, in other words, on tests/assessments as measures of instruction/schooling/education. We are concerned not only with improving assessment of what students know but also with how differences in what they know come about from educational experiences. In other words, we approach testing/assessment from the instruction side rather than the abilities (the nature of knowledge) side.

Adopting our conception would mean that one's model of a student response to a specific test item has to include components having to do with ability attributes and experience attributes, broadly conceived, plus task characteristic attributes and how they interact with the other two types of attributes. How this falls out in a given assessment situation, then, can get quite complex, but some of the issues are clear.

1. A primary purpose for instructionally sensitive assessments is to reflect student knowledge/ability as the consequence of instruction. Whether the specific case is educational effects/school effectiveness research, traditional policy analysis about educational productivity, indicators of educational achievement, or more accountability driven measurement, the quality of the assessment is directly related to the degree that it comprehensively reflects student knowledge/ability at some point in the course of their education/schooling.

2. Within a subject matter, appropriate assessment would invariably require the measurement of multiple dimensions (components) of knowledge/ability. These dimensions/components may entail finer distinctions in representing the topics in subject matter (algebra, geometry) but clearly must also reflect the processes required to perform the specific tasks required by individual test items (e.g., conceptual knowledge, procedural knowledge, a variety of complex thinking skills such as integration of tasks, symbol system flexibility).

3. Each test item reflects multiple components pertinent to the purposes of the assessment. Unfortunately, it also taps components (e.g., reading ability, motivation, test

taking strategy) that are likely to be irrelevant to the judgement about a student's status on the targeted assessment dimensions. Desirable analytical models allow one to assess performance on the desired dimensions and to separate out the irrelevant components.

4. Both content exposure/opportunity to learn and the ways in which students have been taught matter. Instructionally sensitive assessment and analysis technology should detect which students receive more comprehensive exposure and emphases with respect to ability/achievement/knowledge domains of interest. There are "better ways" to teach a topic, but these better ways may depend on individual differences in learning preferences. In the absence of a clear identification of a uniformly best method, instruction that casts the broadest net in introducing ideas is better.

5. Ideally, instructionally sensitive assessments should validly indicate when a particular constellation of educational experiences at the level of interest (student, learning group, class, school, district, state, nation) has achieved the standard of performance judged to be acceptable/essential/good. This conception could refer to a functional literacy standard, an expert standard, or whatever.

6. With respect to any specific outcome domain, the means of presentation of an item and its response format are distinguishable from the underlying attribute. For example, to say that a student "understands" or has "mastered" proportionate reasoning, we mean that regardless of the form of representation of the problem or the mode of response, the student has a high probability of answering a question on the topic correctly.

7. Because students differ in their learning preferences and prior learning experiences, and students (and thus most likely, their learning preferences and prior knowledge/abilities/achievements) are not randomly allocated to/distributed among educational settings, there is a need to distinguish between status/knowledge and progress/learning. Whether educational units should be given "credit" for their students' status/knowledge is unclear (conceivably the units still foster retention or prevent regress) but certainly progress/learning should be properly credited. Developmentalists might be inclined to argue that to credit institutions for the developmental growth is inappropriate but this is really an argument for building a model of "natural development" into one's analysis rather than to discount educational experiences.

Analytically, the implied emphasis on progress/learning means that background attributes of students need to be included in analytical methods (both additively and interactively). If the desired outcomes and their standards are sufficiently comprehensive, however, there should be no problem in crediting institutions for both the proportion of their students who have met the standards and for progress in the direction of achieving standards.

8. Instructionally sensitive psychometrics should distinguish teaching the content domains of interest from "teaching to the test". Specific sensitivity to content irrelevant to the attributes one is trying to measure reflects a form of item bias. In other words, our interest in whether a student knows the specific concept "acute angle" depends on how well knowledge of this concept represents his/her knowledge of the "language" of the properties of triangles. Content exposure effects on specific item performance is undesirable IF one has adequately represented the main dimensions/components reflected in the latent ability/achievement structure. Such effects could reflect "teaching to the specific items contained on an assessment" as opposed to other items that might have been sampled. Approaching the question differently, one wants to discount performance to the degree it reflects specific teaching to the specific item.

Our notion of instructional sensitivity positively values two kinds of opportunity to learn (OTL) effects. The first kind is that general content exposure to a certain topic or of a certain type (e.g., computation, work with word problems) should affect the latent ability/achievement components of interest. This should occur because some teachers teach topics that others don't (topics, not items or narrow concepts) and some teachers emphasize certain ways of teaching topics that others don't.

The second kind of valuable OTL is when the teaching of a specific concept/strategy is particularly synergistic, operating as a "gatekeeper" (once a student understand this idea, the whole picture unfolds). Such ideas/concepts/strategies capture broader understandings and representations that capture broader understandings. In this case, presuming that such key ideas/concepts/strategies existed, opportunity to learn such entities should directly affect corresponding ability/achievement components.

With the first kind of valued OTL, one would want to have effects from content over the set of items representing the topic or method affecting the ability/achievement components which tap these attributes. Some form of compositing improves the measurement properties of the OTL measure here if one can work out the classification system. With the second kind of valued OTL, it is the direct effect of OTL on the gatekeeper idea/concept/strategy on the ability/achievement components that should be detected. There are obviously limits to the reliability of single-item OTL measures, but unless the assessment multiply samples the same idea/concept/strategy, there aren't other good choices.

One possible implication of the above is that item level OTL effects on item parameters (from a Muthen-type IRT model) are likely to be generally useful at some stage to detect whether we have removed item-specific components of performance that are irrelevant to the traits of interest or whether we have adequately modelled all the traits of interest . If a cluster of "similar" items have strong effects, then it may be because of failure to adequately model the pertinent ability/achievement components.

9. Implicit in the previous point is that except in the most narrow of achievement assessments, the psychometric model for outcomes will invariably need to incorporate multiple latent ability/achievement dimensions. Given any reasonable degree of variability in instructional practices and their effectiveness, one would want to detect the specific performance dimensions for which variability occurs rather than a general knowledge/ability trait.

Specific Applications of ISP reasoning

Another possible breakthrough resulting from these joint meetings has been rethinking and elaboration of ideas related to the concept of "instructionally sensitive psychometrics". All along our belief has been that we are trying to develop assessment technicology that will be sensitive to the influences of education/schooling/instruction. The Muthen modeling methodology appears to afford the possibility of analytically elaborating IRT models to incorporate such influences. The problem was that there is literally no prior work besides Muthen's that links IRT estimation with structural equation modeling and thus no precedents for how to interpret the results from an instructional perspective.

The need for elaborating our instructional sensitivity thinking became especially clear with the results reported in the Muthen-Kao-Burstein paper. The items that prove to be instructionally sensitive as used there were primarily beginning concepts in a given content area (e.g., which of the following is an acute angle?) suggesting that gross differences in opportunity to learn a content area within the SIMS sample were being detected. On further examination, however, this interpretation was an oversimplification of the results from the analyses. What was actually occuring was that the items that were being called instructionally sensitive were those that exhibited OTL effects on the specific item, over and above the OTL effects on the latent achievement trait. But this type of direct effect on item performance levels is not necessarily good. For instance, when the student is asked the question about acute angles, the intent is not to measure the student's knowledge of acute angles but instead to infer from his performance on this item to knowledge about the properties of triangles (the knowledge or skill subdomain). If the correct response were given simply because the teacher taught the acute angle concept because it specifically appears on the test, then the instructional sensitivity is not to the domain but to the item. This form of sensitivity from our view is much like item bias. Once exposed, everyone does well on the item; in the absence of exposure, the unexposed group does worse not because they are less capable but because they were never introduced to the concept.

Our current thinking on the matters raised in the previous paragraph is that we need to refine and clarify our conception of instructional sensitivity, on the one hand, and think more carefully about the adequacy of achievement trait modeling, on the other. For instance, if there were separate achievement traits for introductory concepts and more advanced mathematics built into the model, we might expect that the direct effects of item-level OTL on item-specific performance might decline for the concept items while these effects would be manifested on the introductory concepts trait. If so, then one would interpret the model to be instructionally sensitive to weak programs that simply don't cover sufficient content rather than to direct teaching to the items on the test. Eventually, we expect that our elaboration and extension of our conception of instructional sensitivity will lead to a conceptual piece (in addition to the technical work in the Muthen papers, but perhaps drawing on examples from that work for illustrative purposes) intended for the broader educational research and measurement community to stimulate dialogue about the current state of instructional assessment. Right now, the collaboration developing around this work seems quite promising and important.

A major activity has been factor analysis of these same SIMS items. This has been a joint effort including both substantive and more psychometric aspects (see discussion above). In terms of psychometric work, the factor analysis modeling has involved determination of appropriate ways to formulate and analyze a set of items where one dominant factor is present - a general mathematics achievement trait. The aim of the analysis is to try to isolate specific factors corresponding to more narrowly defined sets of items. Regular exploratory analysis cannot find such factors due to the dominance of the general factor. Instead a confirmatory factor analysis structure has to be defined from substantive reasoning and then tested out on the data. The specific psychometric considerations in our case involves the use of tetrachoric correlations for dichotomous variables, the analysis of a set of items that haven't all been administered to all students (rotated form - missing data issue), and how to best utilize the teacher-reported opportunity to learn information in the analysis.

Applicability of IRT to the measurement of a trait confounded by heterogeneous learning circumstances some of the rethinking and new directions in our recent work. In particular, we are trying to develop more sophisticated conceptions of the impact of heterogeneous learning environments (what we mean in part by instructional sensitivity) on performance and operationalize these conceptions in various ways within the psychometric modelling. At the same time, we have already moved away from trying to fit a unidimensional achievement trait within the model by adapting Gustafsson's strategies for hierarchical modelling of general and specific abilities. A considerable portion of the empirical work on SIMS data has been toward developing a conceptually appropriate and empirically adequate multidimensional model for the SIMS items as a first step toward incorporating a multidimensional latent trait model within the overall analytical scheme. This is not a typical IRT model nor is Muthen's approach like other efforts to develop multidimensional IRT models. It is simply too early to tell how far we can go with this effort but in our view, it is clearly worth doing.

The psychometric aim of the analysis is to try to isolate specific factors corresponding to more narrowly defined sets of items in a factor analysis modeling when a dominant factor (a general mathematics achievement trait) is present. Regular exploratory analysis cannot find such factors due to the dominance of the general factor. Instead a confirmatory factor analysis structure has to be defined from substantive reasoning and then tested out on the data. This will be the primary strategy wherein the posited substantive structures will be derived from notions about the consequences for cognitive structures of mathematics performance arising from patterns instructional coverage and emphasis in various mathematics classes.