Educational Evaluation and Policy Analysis
Spring, Vol. 8, No. 1, 1986, pp. 45–60

The Valid Use of Student Performance Measures for Teacher Evaluation

Edward Haertel
Stanford University

Student achievement test scores appear promising as indicators of teacher performance, but their use carries significant risks. Inappropriate tests improperly used may encourage undesirable shifts in curricular focus or poor teaching practices, and may unfairly favor teachers of more able classes. It is often said that standardized achievement test batteries are unsuitable for teacher evaluation, but few systematic alternatives have been suggested. The purposes of this paper are to analyze some problems in using student test scores to evaluate teachers and to propose an achievement-based model for teacher evaluation that is effective, affordable, fair, legally defensible, and politically acceptable. The system is designed only for detecting and documenting poor teacher performance; rewarding excellence in teaching is viewed as a separate problem, and is not addressed in this paper. In addition to pretesting and statistical adjustments for student aptitude differences, the proposed system relies upon attendance data and portfolios of student work to distinguish alternative explanations for poor test scores. While no single set of procedures can eliminate all errors, the proposed system, if carefully implemented, could expose teaching to constructive scrutiny, organize objective information about teaching adequacy, and help to guide its improvement.

With increasing frequency, policymakers seek to improve education by attaching rewards and sanctions to student test performance (Anderson & Pipho, 1984). The attractiveness of tests as tools of public policy reflects increased attention to school and teacher accountability and increased dissatisfaction with policies aimed at improving educational inputs rather than outcomes (Mitchell & Encarnation, 1984). It may also have been encouraged by the publicity accorded changes in national average scores on the Scholastic Aptitude Test (SAT) and other examinations. For whatever reasons, minimum competency testing programs have become commonplace across the 50 states; the Council of Chief State School Officers (1984) has recently endorsed in principle the use of tests permitting state-level achievement comparisons; and in California, school districts can now receive cash awards for improving their average scores in the California Assessment Program (CAP).

I thank Edwin M. Bridges, Lee J. Cronbach, Sanford M. Dornbusch, Linda K. Junker, Conrad G. Katzenmeyer, Carole Perlman, and David E. Wiley for their helpful comments on earlier drafts of this manuscript.

In light of widespread and continuing concern over teacher quality (Bridges, 1984), it is not surprising that the use of student performance data has also been advocated for summative teacher evaluation. Cognitive learning ranks high among the goals of schooling, and student achievement tests promise to reveal directly each teacher's success in bringing it about. California's 1984 omnibus education bill, SB 813, called upon the governing board of each school district to establish standards of expected pupil achievement at each grade level in each area of study and to evaluate teacher competency as it relates to the progress of pupils toward these standards. While the details will vary, similar proposals can be expected in other states.

Despite the superficial attractiveness of student test results for teacher evaluation, their use has not figured prominently in teacher dismissal cases (Bridges, 1984) and has been strenuously opposed by the National Education Association (Quinto & McKenna, 1977). It has been argued that student achievement depends on multiple factors, many of which are out of the teacher's control, and also that published, standardized tests are unlikely to match the learning objectives of a particular teacher at a particular time (Millman, 1981). Moreover, tests can measure only a subset of important learning objectives, and if teachers are rated on their students' attainment of just those outcomes, instruction of unmeasured objectives may be slighted (Elliott & Hall, 1985). For all these reasons, it has become a commonplace that standardized student achievement tests are ill-suited for teacher evaluation.

Some problems with achievement-based teacher evaluation can be overcome by using specially designed tests, and others can be avoided or minimized by appropriate procedures for data collection, analysis, and interpretation. The purposes of this paper are to analyze these problems and to propose a model for achievement-based teacher evaluation that is effective, affordable, fair, legally defensible, and politically acceptable. This proposed evaluation system is intended to address only the problem of assuring a minimum level of teacher competence, not the problem of distinguishing degrees of excellence. Just as a minimum competency test would be unsuitable for selecting scholarship recipients, so the system described in this paper would be unsuitable for determining merit pay increases or selecting mentor teachers.

Inferring teacher competence from test scores requires the isolation of teaching effects from other major influences on student test performance. No teacher evaluation plan will be perfect, but it may be possible to build in enough checks and controls for the most powerful and most plausible of these other influences so that inferences about teaching quality are at least defensible. In the language of test validation, the task is to support an interpretation of student test performance as reflecting teacher competence by providing evidence against plausible rival hypotheses or interpretations (Campbell, 1957; Cronbach, 1971).

Competence, Achievement, and Test Performance

This section describes the determinants of student competence, achievement, and test performance. Achievement is first distinguished from competence. Next, student and then school and community influences on achievement are addressed. Any of these influences might become rival hypotheses, explaining achievement differences among students taught equally well. The discussion next turns to differences between achievement itself and achievement as measured by tests, and further rival hypotheses are discovered.

In the section following this one, some additional cautions are offered concerning standardized testing and teacher performance. These are followed in the third and last major section of the paper by a proposal for a system designed to minimize all these problems and risks. The paper concludes with a brief discussion.

Student Competence Versus Achievement

If teaching competence is to be isolated from other influences on students' knowledge and skills, achievement due to school instruction must first be distinguished from competence attained in other ways. Messick (1984) defines this

distinction between achievement and competence as follows:

> Educational achievement essentially refers to what an individual knows and can do in a specified subject area as a consequence of instruction. Educational achievement should be distinguished from the closely related construct of *competence*, which refers to what an individual knows and can do in a subject area however that knowledge and skill are acquired, whether through instruction or experience or whatever. Like achievement, the term "competence" as used here refers to a continuous variable reflecting various degrees of proficiency rather than to its other common usage as a particular standard of adequate or sufficient performance. In terms of content relevance, . . . educational achievement tests would be referenced to curricula or instructional objectives. In contrast, competency tests would be referenced to specified domains of knowledge and skill regardless of curricula. Confusion arises because many competency tests, especially minimum competency tests, are often evaluated in terms of so-called "curricular validity." This would seem to qualify them as mislabeled achievement tests. On the other hand, many standardized achievement tests are based on conceptions of what a student should know about a subject regardless of how that subject was learned. This would seem to qualify them as mislabeled competency tests. (p. 217)

Messick's terminology may cause some confusion here, because his *student competence* is different from the *teacher competence* also being discussed. Where ambiguity is possible, competence will be qualified as either student competence or teacher competence.

Achievement is defined as whatever students can do as a result of instruction, but the term *instruction* requires clarification. Does it include homework? Does it include self-directed study initiated by the student? How about tutoring by a parent or an older sister or brother? For present purposes, instruction logically refers to whatever the teacher being evaluated is responsible for, but there are degrees of responsibility, and it is often shared. If a teacher informs parents of a student's learning difficulties and they arrange for private tutoring, is the teacher responsible for the student's improvement? Suppose the teacher merely gives the student low marks, the student informs her parents, and they arrange for a tutor? Should teachers be credited with inspiring a student's independent study of school subjects? There is no time to dwell on these difficulties; others lie ahead. Recognizing that some ambiguity remains, it may suffice to define instruction as any learning activity directed by the teacher, including homework.

The question also must be confronted of what knowledge counts as achievement. The math teacher who digresses into lectures on beekeeping may be effective in communicating information, but for purposes of teacher evaluation the learning outcomes will not match those of a colleague who sticks to quadratic equations. In the passage just quoted, Messick confines achievement to "a specified subject area," and states that "achievement tests would be referenced to curricula or instructional objectives." The thorny problems of deciding and then describing what should be taught are mercifully beyond the scope of this paper. In the following discussion, it will be assumed that learning objectives have been specified and communicated to the teacher by the principal, department chair, textbook, or curriculum guide, and that achievement refers to whatever knowledge and skills the teacher has been directed to address.

Student Aptitude and Background Influences on Achievement

Initial competence. Students are likely to differ in their initial levels of knowledge and skills, and the prescribed curriculum may be better matched to the needs of some than others. Suppose, for example, that one teacher is assigned a class in which most students have mastered the prerequisites but have not yet been exposed to the material to be taught, a second teacher is assigned a class in which most students lack the prerequisites for the prescribed curriculum, and a third is assigned a class in which most students have already been exposed to the prescribed material. The first teacher can then present the curriculum as planned and anticipate that a test referenced to

curricula or instructional objectives" will reflect the student's achievement gains. The second and third teachers may either teach the prescribed material or adjust their instruction to the levels of their respective students, but no matter which they do, their students' gains with respect to the prescribed curriculum are unlikely to match those of the first teacher's class.

Individual differences. Even among students with the same initial competence, differential gains are to be expected. A precocious child in the fourth grade and a slow child in the sixth grade may be able to read equally well, but after a year the younger child will probably outperform the older. Students varying in their readiness to profit from instruction are said to differ in *aptitude*. Not only general cognitive abilities, but relevant prior instruction, motivation, and specific interactions of these and other learner characteristics with features of the curriculum and instruction will affect academic growth (Cronbach & Snow, 1977). Aggregation of individual students' achievement to the classroom level will tend to even out the effects of small, idiosyncratic aptitude variations, but across schools and even within schools, classes are unlikely to be equivalent with respect to socioeconomic status, quality of prior schooling, and other major correlates and determinants of aptitude for further learning. While any control for aptitude differences must be imperfect, some adjustment appears essential if an evaluation system is to be fairly applied across teachers of dissimilar students.

Home support. Differential achievement is also to be expected because students enjoy varying levels of out-of-school support for learning. Not only may parental support and expectations influence student motivation and effort, but some parents may share directly in the task of instruction itself, reading with children, for example, or assisting them with homework. Variables such as the number of books in the home, provision of a regular place at home for studying, and even living with both parents have been related to school achievement (Hinckley, Beal, Breglio, Haertel, & Wiley, 1979). For present purposes, all of these factors may be grouped with student aptitudes as influences on students' readiness to profit from the teacher's instruction.

Influences on Teaching Effectiveness Beyond the Teacher's Control

Even if differences in initial student competence and other aptitudes are accounted for, fair evaluation requires that (a) all teachers have comparable teaching and nonteaching responsibilities; (b) they be provided with comparable materials, facilities, and time for instruction; (c) they teach in schools with similar learning climates; (d) they enjoy similar levels of instructional support from other teachers; and (e) they all have received adequate training to teach the prescribed curriculum.

The importance of the teaching loads and of other demands on teachers' time should be clear. If class sizes differ substantially, if some teachers have more classes to teach than others, or if some have much less scheduled preparation time, they may be unable to devote comparable effort to preparing for class, writing comments on student papers, working with students individually, or contacting parents. The same is true if some teachers are formally charged with more nonacademic responsibilities than others.

The quality of instructional materials and facilities and the time available for instruction also may influence student achievement (Wiley & Harnischfeger, 1974). Materials and facilities may include the classroom, desks and chalkboards, textbooks, or audiovisual materials, as well as computers or other specialized equipment. Available instructional time may be a function of the length of the school day or instructional period, of time for lunch, recess, and the like, and even of school policies on classroom interruptions (Goodlad, 1984, pp. 95–107).

Schoolwide learning climate refers to the host of factors that make a school more than a collection of self-contained classrooms. Where the principal is a strong instructional leader; where schoolwide policies on attendance, drug use, and discipline are consistently enforced; where the dominant peer culture is achievement-oriented; and where the school is actively supported by parents and the community, individual teachers

can be more effective than in schools with less favorable learning climates (Bridge, Judd, & Moock, 1979; Brookover, Beady, Flood, Schweitzer, & Wisenbaker, 1979; Goodlad, 1984; Wellisch, MacQueen, Carriere, & Duck, 1978).

Instructional support from other teachers takes a variety of forms. In elementary schools, the classroom teacher may be assisted by resource teachers and aides, some students may leave the classroom for pull-out programs, or selected subjects may be team-taught. In high schools, reading and writing may be directly taught in English classes, but they are practiced almost everywhere, and the English teacher's task will be more or less difficult depending on the amount of concurrent practice given by other teachers and the quality they demand (Applebee et al., 1984). The physics teacher can reinforce the work of the algebra teacher, and the Latin teacher can help the biology teacher.

Teacher responsibilities, teaching facilities, school climate, and support from other faculty are largely beyond the control of most individual teachers. There should be little disagreement that two teachers for whom these factors differ grossly should not be held to the same standard of student achievement. The last of the five teaching effectiveness resources, adequate teacher training for the prescribed curriculum, appears different. A physician or attorney who made some serious error in a given case could scarcely plead poor training as an excuse. The teacher, however, especially at the secondary school level, has less control than the doctor or lawyer over which cases, or curricula, to accept. If funding cutbacks force the assignment of a school psychologist to teach special education classes, or if pressures to increase science enrollments in the face of a science teacher shortage lead to the assignment of home economics or physical education teachers to biology classes, their only alternatives may be to accept the challenge or to resign their jobs. It appears unfair to compare student achievement for a teacher in this situation to that for a teacher properly trained. A similar argument might apply to inexperienced teachers during their first year or two in the classroom, or even to teachers required to adopt a curriculum radically different from the one to which they are accustomed.

Achievement Versus Test Performance

The most practical tools for quantifying student achievement objectively are written examinations. It is not achievement itself but student test performance that must be linked to teacher competence. The best of tests would be imperfect, and widely used achievement tests are not the best that could be designed for evaluating teacher competence. Thus, measurement effects must also be taken into account in designing a teacher evaluation system.

Measurement using an ideal test. All that any test provides is a sample of student performance. The inference that this performance reflects educational achievement is probabilistic, and is only justified under certain conditions. Obviously, the content tested must match the curriculum or instructional objectives for which the teacher is responsible, but even a perfect match would not assure that test performance was attributable to instruction by the teacher being evaluated. Important topics are likely to have been presented earlier in the school curriculum, and so test scores may reflect the instruction of earlier years. Students may also have acquired tested knowledge and skills outside school. It might be safe to assume that initial competence was uniformly low in the first year of an unusual foreign language, but in most courses and content areas some control for initial student competence would be required. For this and other reasons, Messick (1984) argues that the measurement of educational achievement must employ observations at multiple points in time. Preinstruction to postinstruction changes on well-constructed tests matched to the curriculum may generally be attributed in large part to the intervening instruction.

Of course, even with content valid tests pretest to posttest gains may have alternative explanations. Gains might reflect no more than test taking practice, especially if the examinees are young children or if the test employs an unfamiliar format. If students believe that they are to be evaluated on the basis of their gains,

or that their pretest performance will influence the pace or level of instruction they are to receive, they may not do their best on the initial examination. Teachers might vary testing conditions from pretest to posttest, either accidentally or intentionally. Finally, if the same form of a test is used repeatedly or if test security is breached, students may learn the answers to specific items, destroying their validity as measures of any broader outcome domain.

Measurement using standardized tests. The standardized tests referred to here are objective, paper-and-pencil instruments, group administered, and, with the exception of writing assessments, usually machine scored. Their administration may be mandated by a state or school district, and individual teachers generally have little control over their selection or use. Districtwide testing programs account for most standardized testing, but statewide or nationwide programs like CAP or the National Assessment of Educational Progress also fall in this category.

The match of a standardized test to instruction at any particular time in any particular classroom is likely to be poor (Millman, 1981). The logistics of their marketing and administration have led to the design of instruments more suitable for measuring the cumulative effects of years of schooling than the effects of instruction over time intervals as brief as a semester or less. Tests must be broad in their content coverage if they are to be used across schools and classrooms with varying curricula, varying instructional pacing, and texts that may sequence material differently; tests closely tied to the curriculum would have to be tailored to virtually every school system and carefully synchronized with instruction.

As their breadth of focus increases, tests become less suitable as measures of teaching quality for two reasons. First, test scores become more sensitive to student individual differences beyond the teacher's control, including initial differences in learning aptitude due to the quality of instruction in years past. Second, test scores become less sensitive to the quality of current instruction. A test on the content of a brief series of lessons, given shortly after they are taught, might provide a fair assessment of the quality of the teaching. The broader the content of the test, and the less closely tied to the specific content of the teacher's own curriculum, the less fair that assessment becomes (Haertel, 1985).

Summary

Inferring teacher competence from student test performance requires that test score determinants other than teaching quality be accounted for. Educational achievement, due to instruction, must be distinguished from student competence, which may be acquired in other settings. Influences other than teaching on amount of educational achievement must be controlled. Besides initial student competence, factors to be considered include student aptitudes and background characteristics (e.g., general mental ability, academic motivation, and parental support); class size and teaching load; instructional time, materials, and facilities; school climate; concurrent instruction by other teachers; and in some cases, teacher training and experience.

The measurement of teacher competence may also be distorted by testing artifacts. Test content must match the prescribed curriculum; students must be motivated to do their best on both pretests and posttests; familiar testing procedures and test formats must be used to minimize practice effects; consistent testing conditions must be assured; and the tests themselves must be kept secure.

The use of standardized achievement test results for teacher evaluation carries special risks. Standardized tests tend to be unsuitable for measuring educational achievement as distinct from student competence, because they sample broad subject domains and are unlikely to match closely the curriculum in particular classrooms at particular times. Their breadth of focus makes such tests more sensitive to student individual differences beyond the teacher's control and less sensitive to the quality of current instruction.

If extraneous factors are either held constant across teachers or controlled statistically, and if suitable tests are used, then a substantial fraction of the remaining achievement variation across class-

rooms may be attributable to differences in teacher competence.

Using Standardized Tests for Teacher Evaluation

Administrators may value the information standardized tests provide, but it appears to have little relevance to students and teachers. When asked about testing, high school students discuss the teacher-made or teacher-selected classroom tests on which their grades depend, and rarely comment on standardized tests. Even the minimum competency tests they must pass for high school graduation are not a serious issue for most students (Haertel, Ferrara, Korpi, & Prescott, 1984). Teachers rarely consult standardized test results except, perhaps, for initial grouping or placement of students, and they believe that the tests are of more value to school or district administrators than to themselves (Herman & Dorr-Bremme, 1983). Moreover, teachers report that observations and ratings of student behavior are as important or more important than test scores for their instructional decisionmaking in many subject areas (Stiggins & Bridgeford, 1984).

If standardized tests are remote from the day-to-day concerns of the classroom, then as policy tools they must at best be blunt instruments, not closely tied to any of the particular teacher behaviors and instructional activities through which teaching and learning occur. Policymakers might understand this, however, and still choose to attach sanctions to test scores. They might be unconcerned with the precise mechanisms by which their policies influenced teacher or administrator behavior, acting instead on the belief that if teachers simply exerted more effort, achievement would increase. Alternatively, they might believe that excessive instructional time and resources were being devoted to nonacademic objectives, and that testing would encourage an increased allocation of resources to teaching and learning the tested skills.

Incentives To Alter Curriculum and Instruction

Teachers can indeed alter instruction in an attempt to improve test scores, but it is not necessarily desirable that they do so. The measures taken by high schools to improve SAT scores illustrate one possible response when rewards and sanctions are attached to test performance. Because the SAT is important to students, parents, and teachers, coaching classes have appeared in high school curricula, in which students study basic test-taking strategies, such as intelligent guessing, pacing, deferring difficult problems, and reviewing their answers; and practice taking items like those they will encounter on the SAT. Direct instruction of this kind is probably more effective in raising high school students' SAT scores than any other instructional intervention of comparable cost and scope (Messick, 1980), but it is doubtful that the coaching has many long-term benefits once that test has been taken. Rewarding teachers for their pupils' standardized test scores invites even more instruction focused on test performance and little else.

In addition to encouraging direct instruction in test-taking skills, rewarding teachers, schools, or districts for good test scores may lead to a gradual narrowing of the curriculum to just the knowledge and skills tested (Elliott & Hall, 1985; Frederiksen, 1984). While this might appear at first to be what policymakers intend, it must be recognized that the restriction of curricular focus would have content and process aspects. It is obvious that if rewards or sanctions depend primarily on test scores, teachers will have little incentive to devote instructional time to *content* that they know will not be tested. Less obvious but equally important, teachers will have little incentive to have students engage in *instructional activities* that don't resemble the tests in format.

It is not suggested here that a teacher evaluation plan involving student test performance would lead immediately to massive drops in the amount of class discussion, extended out-of-class writing or other projects, any more than it would precipitate sweeping changes to focus the curriculum on just what was assessed. A more likely consequence would be a gradual shift toward more in-class objective testing, both to give students practice with tests and to alert the teacher to areas of weakness in test performance; a shift toward fill-in and short-answer or multiple-

choice exercises on worksheets, away from less structured activities; and changes in content emphasis that over time would lead to a closer match of classroom instruction to the pattern of content coverage on the test. None of these changes would be unethical or in themselves obviously poor pedagogy, but taken together, they could easily result in students being less well prepared to apply their academic skills outside the classroom.

Limitations of Objective Tests

It might appear that this problem could be solved with better tests. If all important learning outcomes were assessed, and if students were required to demonstrate their content mastery in appropriate ways, then the incentives created by testing would be entirely salutary. Unfortunately, the state of the art in measurement is such that reliable, valid, and efficient measurement may not be possible for many important cognitive learning outcomes (Frederiksen, 1984; Haertel, 1985; Messick, 1984; Stiggins & Bridgeford, 1984). The problem is not simply one of lower order objectives, such as rote memorization, being easy to test and higher level objectives, such as analysis, synthesis or evaluation, being difficult to measure. It arises, rather, from dissimilarities between the activity of test taking and the range of situations in which school learning ought to matter.

Present testing technology is such that efficient, objective measurement is possible only with fixed-alternative formats, for example, multiple-choice items. Regardless of the item writer's cleverness in stimulating careful analysis or problem solving, a multiple-choice item remains a recognition task, in which the problem is to find the best of a small number of predetermined alternatives and the criteria for comparing the alternatives are well defined. The nonacademic situations where school learning is ultimately applied rarely present problems in this neat, closed form. Discovery and definition of the problem itself and production of a variety of solutions are called for, not selection among a set of fixed alternatives. The solutions found are not right or wrong, but more or less adequate along a variety of dimensions (Frederiksen, 1984). It is not likely that instruction aimed at improving test scores is also optimal for improving performance in such situations (Haertel, 1985). In fact, the relationship between test performance and out-of-school performance is surprisingly weak. Except for their ability to predict success in further schooling, the predictive power of academic achievement tests has been disappointing (McClelland, 1973).

Summary

An attempt to alter educational practices by testing intended learning outcomes and attaching sanctions to test performance may result in higher test scores without improving teaching or learning. While such a policy can encourage a desirable shift in curricular content toward the material tested, a less beneficial, concomitant shift in instructional methods is possible, toward more testing and activities resembling tests. In addition, it is likely that not all important instructional outcomes will be assessed, and that basing rewards on measures of some will lead to decreased emphasis on the others. Some of these difficulties with standardized tests could be overcome through better test design, but others represent fundamental limitations of objective, paper-and-pencil measures.

A Proposal for Teacher Evaluation Based on Student Achievement

If teacher evaluation based on student achievement is to be fair across instructors and if incentives for undesirable teaching practices are to be minimized, careful attention must be paid to the formation of appropriate teacher comparison groups, to test design and data collection, to data analysis and interpretation, and to the problem of setting performance standards. These areas are addressed below. In addition to achievement tests, student achievement portfolios are described, which it would be the teacher's responsibility to assemble. Also required would be data on individual student attendance, either daily or on a random 10% or more of school days, and the average chronological age of the students in each classroom.

Initial implementation of the evalua-

tion system might focus on a single major learning objective and a range of several grade levels during which growth on that objective was expected. After this system was in place, it could be expanded to cover additional grade levels and learning outcomes. For ease of exposition, the following discussion is phrased in terms of a single learning objective. This might be reading comprehension, writing, or mathematics problem solving as taught in the upper elementary grades.

Expansion beyond the initial system would entail no new fundamental problems, although various technical difficulties might arise. Vocational courses would probably require applied performance tests rather than paper-and-pencil measures, and in areas where outcomes cannot be well represented by a single, unidimensional continuum of skill acquisition, multiple tests or subtests might be required. For example, different approaches to the teaching of a foreign language could result in various profiles of oral, aural, reading, and writing competencies, and in different balances of linguistic versus communicative competence. Several types of test items would be required to cover this range of learning outcomes, and decisions about the relative number or weight of items of each kind would implicitly define the relative importance of the outcomes.

Appropriate Teacher Comparison Groups

Ideally, it might be possible to use theories of curriculum, instruction, measurement, and psychology to quantify the amount of achievement a particular student ought to attain under competent instruction. In practice, the only workable approach would be to use the achievement of students in similar classrooms as a guide in setting expectations. Teacher evaluation based on student achievement must be norm-referenced for the foreseeable future. The problem, then, is to define appropriate norming or comparison groups, to determine which other classrooms should be considered similar to any given classroom. In the proposed evaluation system, similar classrooms would be those in which teachers were directed to address the same learning objectives, taught comparable students, and had access to comparable school resources.

The requirement that teachers be responsible for addressing the same objectives refers not only to the objectives in the content area evaluated, but in a rough way to other content areas as well. A major effort by one school or district to improve mathematics achievement could place its teachers at a disadvantage relative to others on a test of reading comprehension, by diverting instructional time from reading to math. Comparisons would have to be restricted to teachers at the same grade level, and where tracking or streaming was practiced, to single tracks. Not only is the pace of instruction likely to increase with track or grade level, but in most skill areas, different component processes may be emphasized. In reading, for example, letter-sound correspondences and decoding skills are the focus of instruction at the primary level, while comprehension skills receive increasing emphasis in the upper elementary grades (Calfee & Drum, 1978). Students whose skill levels match the instruction they receive are likely to show the largest gains, so that different ranges of pretest scores may be associated with maximum growth, depending on track and grade.

The requirements for comparable students and resources imply that comparison groups would have to be restricted to similar schools and communities as well. While statistical controls are proposed for differences in student aptitude and initial competence, these rough adjustments probably would not permit equitable comparisons across areas differing greatly in socioeconomic status, proportion of bilingual students, urbanization, or other demographic characteristics strongly related to academic achievement.

When classrooms are divided into homogeneous groups to establish norms, there is a tradeoff between bias and precision. Increasing the number and reducing the size of the comparison groups should yield norms with the smaller biases, because each classroom is compared to a sample of classrooms more like itself. At the same time, as the size of each group diminishes, the precision with which its characteristics can be estimated also diminishes. A minimum comparison

group size of at least a few dozen classrooms probably would be required. The evaluation system might best be implemented in single large school districts or in groups of smaller districts in the same area that had agreed to a common set of curricular objectives.

Even within a single track and grade level across a group of similar schools, it probably would be unfair to compare certain individual teachers to their peers. Teachers might be exempted from an annual evaluation if they had been teaching only 1 or 2 years, had recently changed curriculum areas or grade levels, or were charged with unusually heavy teaching or nonteaching responsibilities.

Test Design and Data Collection

Test scales to measure student progress. For general skills, such as reading comprehension, writing, or mathematics problem solving, items could be written so that even instruction aimed only at improving test performance would also develop the general skill. This is desirable to minimize the negative impact of the testing requirement on curriculum and instruction. It would be accomplished by writing items that required application of knowledge rather than factual recall, by measuring each skill using several item formats, and by relying where possible on free-response items with objective scoring guides rather than on multiple-choice items.

As discussed in the last section, pretesting would be required in most courses to help distinguish educational achievement from competence attained in other settings. In addition, if teachers were to be held accountable for student progress on specific tests, it would be desirable to provide practice forms of the tests for them to use at their own discretion and to inform them of the end-of-year performance levels expected.

The multiple test forms required for pretesting, practice, and posttesting could be assembled and equated most easily from a pool of items calibrated using Item Response Theory (IRT). Unlike classical test theory, which models scores on entire tests, the basic measuring unit in IRT is a single item. The probability of a correct response to each item is modeled as a function of one or more parameters representing examinee abilities and one or more parameters unique to that item. These item parameters are estimated in a process referred to as item calibration, and are generally assumed to reflect invariant properties of the items (however, see Bock, Cook, & Pfeiffenberger, 1985; Goldstein, 1979). Once a set of items has been calibrated, any subset of them can be used, theoretically, to estimate an examinee's ability on the same score scale as any other subset (Lord, 1980).

Given a calibrated pool of items measuring the target skill, some could be reserved for secure pretests and posttests, while others could be formed into practice tests, examined by teachers, published to illustrate the abilities measured, or used for classroom instruction. Scores on all of the tests could be reported on a common scale, anchored by descriptions of the capabilities of students scoring at different levels and by examples of items on which those students had a specified probability of success (Bock, Mislevy, & Woodson, 1982). Simple tables could be constructed permitting teachers to convert number-correct scores on each practice test to the common scale.[1]

When number-correct scores are used, smaller pretest to posttest gains may be expected for students initially near the chance level (floor) or highest possible score (ceiling) of a test than for those beginning near the middle of the score range. In the proposed system, floor and ceiling effects could penalize teachers for whose classes the test was of inappropriate difficulty. These effects could be minimized by using a test with a wide range of item difficulties, but such a test might have to be prohibitively long to provide sufficient reliability. IRT score re-

[1] It is sometimes stated that using number-correct scores implies the use of the Rasch model rather than more complex IRT models (e.g., Wright & Stone, 1979), but construction of raw score to scaled score conversion tables does not necessarily require use of a Rasch model. It is true that under other IRT models, better ability estimates can be calculated from more complex scorings than from number-correct scores, but unbiased ability estimates corresponding to all but the extreme number-correct scores are readily obtained under any of the common IRT models, and would be sufficiently accurate for instructional decisionmaking.

porting rather than number-correct score reporting would permit the use of focused, reliable tests at different levels of difficulty and reporting of scores on all of these tests on a common scale.

Achievement test data collection. All students would be pretested near the beginning of the academic year and posttested at the year's end. While no test administration procedures can guarantee against problems of cheating or nonuniform testing conditions, other examinations of comparable importance are administered successfully, and it should be possible with reasonable care to secure reliable data. Efforts would be made to secure both pretest and posttest scores from all students, including those absent on the regular testing days. However, students joining a class after the pretesting or no longer enrolled at the time of posttesting would not be included. Exclusion of these students is recommended because their previous educational experience is uncontrolled and because they do not receive a full year of instruction from the teacher being evaluated. It is conceivable that this provision would create an incentive for marginal teachers to encourage dropout or transfer of weak students. If this proved to be a significant problem, concomitant monitoring of dropouts and transfers might be required.

If students believe that they are to be evaluated on the basis of change scores or that their pretest performance will influence the level of instruction they are to receive, or if their teacher is unpopular, they may not exert maximum effort. To help assure uniformly high motivation, students' scores could be sent to their parents after each testing. Pretest scores would accompany reports of posttest performance, and when available, students' posttest scores from the previous year could be sent along with their pretest scores. While students still might have an incentive to do poorly on their initial pretests, any such tendency would bias the evaluation in favor of the teacher. A few teachers also might attempt to influence pretest scores (negatively) or posttest scores (positively). Outside proctors could be brought in to administer these examinations, although better teacher-administrator relations probably would be maintained if external proctoring were avoided.

Student achievement portfolios. Teachers would be individually responsible for assembling portfolios of each student's work. These might include completed practice tests, regular classroom tests used at the discretion of the teacher, samples of student themes or other written work, occasional examples of homework papers, and, especially at the lower grade levels, notes on the teacher's observations of individual students. For students not progressing satisfactorily, it would be the teacher's responsibility to use the portfolio to document attempts at remediation. Records might be included of individual conferences with the student or parents, as well as copies of requests for consultations by resource teachers, or of extra, remedial work assigned. For uncooperative students, it would at least be possible to include copies of invitations to the student to meet individually with the teacher or letters to the student's parents.

The student portfolios would serve several purposes in the proposed evaluation system by providing some context in which to interpret achievement data (Messick, 1984). They would supplement grade books with different kinds of actual work samples, aiding teachers in monitoring individual student progress. If a student who had done well during the year failed to show satisfactory progress on the posttest, the teacher could use the completed practice tests to document that the posttest performance was anomalous. Regardless of students' test performance, teacher evaluation would include inspection of the portfolios to provide some minimal evidence that an acceptable variety of instructional techniques was employed, and that instruction was provided for learning outcomes not evaluated formally.

Data Analysis and Interpretation

Overview. Pretest, posttest, attendance, and achievement portfolio data would be examined systematically, following a procedure designed to rule out as many alternative explanations as possible before holding any teacher responsible for poor student performance. Prior to implementation, pilot studies within each teacher

comparison group would be conducted to establish norms for typical growth over an academic year. Using multiple regression, average classroom growth would be predicted as a function of average student age and average pretest score for a classroom. Residualized gain scores would be computed by subtracting predicted growth from observed growth.

Through a process of informed deliberation, minimum standards would be established for these residualized gains, and these standards would be used to determine for each teacher whether overall student progress was satisfactory. However, a finding of unsatisfactory overall progress would only trigger a detailed student-level examination involving attendance data and student portfolios. A teacher would not be put on notice or otherwise sanctioned unless this second, detailed examination proved unsatisfactory.

In the detailed examination, some students would be exempted for any of several causes, and the mean growth of those remaining would be judged against a separate standard, somewhat higher than the initial one. A teacher would not be held accountable if a given student's portfolio gave evidence of satisfactory progress in spite of an unexpectedly poor posttest score, if the student's attendance was poor, or if the portfolio showed evidence that learning difficulties had been recognized and that special remediation had been attempted. A teacher would fail only if posttest performance was unsatisfactory for students who had failed to show satisfactory progress during the year, attended class regularly, and received no special assistance from the teacher. The steps of this procedure are described in greater detail and justified in the remainder of this section.

Pilot studies to establish norms for classroom gains. For each comparison group, a separate pilot study would be conducted. These studies probably would use essentially the same teachers as were to be evaluated. After an initial year of pilot studies, a second year would be devoted to a trial implementation, and actual teacher evaluation would begin in the third year.

Each pilot study would involve collection of pretest, posttest, attendance, and portfolio data, as well as student chronological age. The age and test data would be analyzed using multiple regression with the classroom as the unit of analysis. Mean classroom posttest score would be modeled as a function of mean pretest score and mean chronological age. Fitted regressions would be smooth, but not necessarily linear. Portfolio and attendance data would be used in simulations of the entire evaluation procedure during the standard-setting process, as described below.

Pretest score and chronological age are proposed as predictors in the regressions for each comparison group to provide a partial control for initial student competence, reflected in pretest scores, and for aptitude, which should be inversely related to chronological age at any given pretest level. In general, these variables should not be difficult to obtain for most students. Classroom-level rather than student-level regressions are proposed because the alternative of predicting gains for each student and then aggregating to the classroom level would be biased in favor of some teachers and against others. This is because the classroom-level regression, pooled within-classroom regression, and total regression need not coincide (e.g., see Burstein, 1978).

The pilot studies would provide regression equations to be used in calculating each classroom's predicted growth, and when the evaluation system was implemented, predicted growth would be subtracted from observed growth to obtain each classroom's residualized gain score.

Procedure for teacher evaluation. The residualized gain score for each classroom would be compared to a single standard and judged satisfactory or unsatisfactory. (The problem of setting standards for these judgments is discussed below.) If this initial comparison indicated satisfactory progress, no further review would be required, although the teacher's student portfolios would be examined briefly to see what kinds of work the teacher had required and what students had accomplished over the year. If the initial step indicated unsatisfactory progress, attendance data and portfolios would be reviewed for all students in the class. Stu-

dents meeting specified criteria would be excluded from the analysis, and the predicted growth, observed growth, and residualized gain score would be recalculated using only the remaining students. The revised gain score would be judged against a more stringent standard than was used initially.

Grounds for excluding students would be clearly defined, but their detailed specification would require careful deliberation, tryout, and probably revision. It is proposed here that three groups of students would be excluded. The first two groups would be students absent more than a certain number of days and those whose posttest performance was markedly poorer than their performance on a series of earlier practice tests. The third group would include students who performed poorly despite special efforts by the teacher to assist them. These would be students with low gain scores for whom teachers could document that (a) unsatisfactory progress had been noted prior to the posttest, (b) parents had been informed, and (c) appropriate remediation had been attempted. This might include individual student conferences or parent conferences with the classroom teacher, special homework assignments, or consultations with resource teachers. Students could also be excluded if such special assistance had been offered and was refused.

Attendant risks of the proposed system. The simplest teacher evaluation system using student test performance might be to administer a standardized test at the end of the year and to place teachers on notice if their students averaged, say, more than 1 year below grade level. Such a system would be unfair to many teachers and could encourage curricular distortions and undesirable teaching methods. The proposed system is designed to minimize adverse impacts, but some remain. As already noted, pressures could be created to narrow the range of content covered and teaching methods employed, and marginal teachers might encourage weak students to drop out or transfer to other classes. The system could create teacher morale problems, especially if pretests and posttests were given by outside examiners. Two additional cautions are in order. The first concerns allocations of instructional resources among students in a classroom, and the second concerns the creation of incentives for poor student attendance.

In the system described, teachers are judged on the basis of their classes' average growth. To maximize average growth, a teacher's best strategy would be to invest the most time in those students progressing fastest, regardless of their skill levels. These probably would be the students who had made the most rapid progress in the past, so that teachers might be discouraged from spending the extra time required to help slower learners. Provisions in the proposed system for excluding low-scoring students who had been given special help might only partially counteract this incentive.

The last difficulty to be discussed concerns incentives for poor student attendance. In the proposed system, teachers are not held accountable for the achievement of students absent more than some specified number of days. This provision could conceivably lead teachers to encourage poor attendance. More important, research has shown that poor student attendance at the high school level may be a result of poor teaching. Students are more likely to cut classes in which little content is covered and performance expectations are low (Natriello & Dornbusch, 1984). Thus, while it may be appropriate not to hold teachers accountable for the performance of students who are often absent, it must be noted that excessive absenteeism can be symptomatic of poor teaching. Teaching might be monitored in high school classes where excessive cutting was found, independent of the proposed evaluation system.

Setting Standards

Up to this point, a largely technical procedure has been described. Significant value questions would arise in selecting the outcomes to be assessed, writing items, and establishing guidelines for the construction of portfolios, but the central problem of setting performance standards for teacher evaluation has not been addressed. Establishment of these standards is considered in this section.

Criteria for defensible standards. The

one purpose of the proposed evaluation system would be to assure a minimum level of teacher competence, and this would have to be clear to all participants in the standard-setting process. The goals of detecting incompetence and of rewarding excellence require qualitatively different kinds of information, and any evaluation system that attempted both would risk doing neither well.

Ideally, then, standards should be high enough that incompetent teachers would fail and marginal teachers would be pressed to exert more effort, yet low enough that competent and effective teachers would not need to distort their teaching to improve test scores. Determining such standards, if it is technically and politically possible, would require informed deliberation by representatives of teachers, administrators, and the public, aided by simulations based on pilot data, and confirmed by a trial implementation period during which no penalties were imposed for teacher failure.

Overview. The standards established must represent some meaningful level of academic achievement, and in this sense they are criterion-referenced. At the same time, they must reflect typical levels of classroom performance, and in this sense they are norm-referenced. Accordingly, participants in setting standards would be acquainted with the IRT score scale through descriptions of the abilities represented by different score levels and through illustrations of items students at different levels should be able to answer. They would also be informed of classroom performance norms based on the pilot data. As tentative standards were formulated, participants would be given projections of the numbers of teachers likely to fail under those standards, again based on the pilot data. After initial standards were agreed upon, their impacts would be monitored through a 1-year trial implementation period and revised if necessary. Following implementation of the evaluation system, standards and procedures would be monitored, reviewed, and possibly revised annually.

For each teacher comparison group, standards would take the form of two numerical values: (a) the minimum residualized gain score required to pass the evaluation without a detailed, student-level review, and (b) a second, higher score to be used in judging the residualized gain for nonexcluded students after a detailed review. Each comparison group's numerical values would be established independently of the others, but for all groups a common set of procedures would be followed. The first of the two numerical values would be established through a process of informed deliberation, simulation based on pilot data, trial implementation, and revision. Derivation of the second, higher value would require no further judgments.

In addition to these two numerical criteria, detailed guidelines would be established for excluding students showing anomalously poor posttest performance following satisfactory work during the year, students with poor attendance records, and students for whom learning difficulties were recognized and remediation was attempted. These guidelines would be as uniform as possible across comparison groups, although with some inevitable differences across grade levels and subject areas. They would be established concurrently with the numerical values and also would be subject to tryout, revision, and annual review.

Minimum residualized gain scores to bypass student-level review. Residualized gain scores would indicate the relation between actual and expected growth in each classroom. Their mean would be close to zero, and if classrooms were sampled from a homogeneous population, they would probably be approximately normally distributed, with a standard deviation slightly larger than the standard error of estimate from the comparison group's pilot regression. It is likely that the distribution of residualized gain scores would have heavy tails, with larger numbers of extreme positive and extreme negative values than normal theory would suggest. An initial standard of negative two standard errors, for example, would probably trigger detailed reviews for somewhat more than 2.2% of the teachers.

The appropriate standard would depend on the shape and the dispersion of the distribution of the residualized gains and on judgments of the number of sub-

standard teachers in the system. If the residualized gains for a handful of poor teachers were clear outliers, a standard might be established to separate these values from the rest of the distribution. If there were no clear outliers, a standard might be established to fail some specified proportion of the teachers, probably somewhat less than the estimated proportion of substandard teachers.

Minimum gain scores for nonexcluded students following review. After rules had been established for excluding students based on poor attendance, inconsistent performance, or special learning difficulties, pilot data would be used to calculate the second, higher criterion value for each teacher comparison group. The procedure essentially would be to predict the gain for nonexcluded students from the gain for all students, using only classrooms in which teacher performance appeared satisfactory. This prediction would then be applied to the original cutting score to obtain the second, higher cutting score. Details of the procedure are as follows.

First, any classrooms in the pilot sample with residualized gains below the initial criterion value would be set aside and would not be used further in the analysis. The average residualized gain for the remaining classrooms would be calculated. If any classrooms were excluded, this mean would be positive; otherwise, it would be exactly zero.

Next, for each of the remaining classrooms, all students' portfolios and attendance data would be reviewed to determine which students met any of the exclusion criteria, and the observed gain for the students remaining would be calculated. Using the original regression equation, predicted gain scores would also be calculated for these remaining students, and the difference between observed and predicted values would be taken to obtain a second residualized gain score for the classroom, based only on nonexcluded students.

Finally, the mean across classrooms of these new residualized gains would be calculated. Assuming that the students excluded were generally the poorer performers, this second mean would be higher than the average calculated earlier of the original residualized gains. The original average would be subtracted from the new average, and this difference would be added to the initial criterion value to obtain the second, higher criterion value for nonexcluded students.

Conclusion

A system has been proposed for teacher evaluation based on student performance data. Logical, psychometric, and statistical problems have been examined, and some solutions have been suggested.

Compared to using off-the-shelf, standardized tests, the system proposed appears expensive and complicated. Special tests are recommended, formed from a bank of IRT-calibrated items; pretesting, posttesting, and practice testing at teachers' discretion are called for; and moderately complicated statistical procedures are required. Teachers would be expected to keep records and samples of each student's work, to be inspected routinely and to be used, if necessary, to justify unsatisfactory student progress. Clearly, this evaluation system would be burdensome, and most school districts would require expert psychometric and statistical assistance to carry it out. Implementation would not be quick, easy, or inexpensive.

In presenting the proposal, an attempt has been made to justify each component, each complication. It has been argued that the refinements proposed are necessary to safeguard teachers' rights and to minimize incentives to poor pedagogical practice. The greatest bulk of the costs would be incurred in the initial implementation of the system; the IRT methods recommended and the objective procedures to be specified should lead to lower costs in the long run. No measurement or evaluation procedure can eliminate errors of measurement or classification, and all systems can be circumvented, but the proposed procedures could expose teacher effectiveness to constructive scrutiny, organize and summarize objective judgments of its adequacy, and guide improvement.

References

ANDERSON, B., & PIPHO, C. (1984). State-mandated testing and the fate of local control. *Phi Delta Kappan*, 66, 209–212.

APPLEBEE, A. N., LANGER, J. A., DURST, R. K., BUTLER-

NALIN, K., MARSHALL, J. D., & NEWELL, G. E. (1984). *Contexts for learning to write.* Norwood, NJ: Ablex.

BOCK, R. D., COOK, L., & PFEIFFENBERGER, W. (1985, June). *Accounting for item parameter drift.* Paper presented at the meeting of the Psychometric Society, Nashville, TN.

BOCK, R. D., MISLEVY, R., & WOODSON, C. (1982). The next stage in educational assessment. *Educational Researcher, 11*(3), 4–11, 16.

BRIDGE, R. G., JUDD, C. M., & MOOCK, P. R. (1979). *The determinants of educational outcomes: The impact of families, peers, teachers, and schools.* Cambridge, MA: Ballinger.

BRIDGES, E. M. (with B. GROVES). (1984). *Managing the incompetent teacher* (ERIC/CEM School Management Digest Series, Number 29). Eugene, OR: University of Oregon, College of Education, ERIC Clearinghouse on Educational Management. (ERIC/CEM Accession No. EA 016 645).

BROOKOVER, W., BEADY, C., FLOOD, P., SCHWEITZER, J., & WISENBAKER, J. (1979). *School social systems and student achievement: Schools can make a difference.* New York: Praeger.

BURSTEIN, L. (1978). Assessing differences between grouped and individual-level regression coefficients. *Sociological Methods and Research, 7,* 5–28.

CALFEE, R. C., & DRUM, P. A. (1978). Learning to read: Theory, research, and practice. *Curriculum Inquiry, 8,* 183–249.

CAMPBELL, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54,* 297–313.

Council of Chief State School Officers. (1984, November). *Education evaluation and assessment in the United States. Position paper and recommendations for action.* (Available from Council of Chief State School Officers, 400 North Capitol Street NW, Suite 379, Washington, DC 20001)

CRONBACH, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.), pp. 443–507. Washington, DC: American Council on Education.

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

ELLIOTT, E. J., & HALL, R. (1985). Indicators of performance: Measuring the educators. *Educational Measurement: Issues and Practice, 4*(2), 6–9.

FREDERIKSEN, N. (1984). The real test bias. Influences of testing on teaching and learning. *American Psychologist, 39,* 193–202.

GOLDSTEIN, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal, 5,* 211–220.

GOODLAD, J. I. (1984). *A place called school.* New York: McGraw-Hill.

HAERTEL, E. H. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research, 55,* 23–46.

HAERTEL, E. H., FERRARA, S., KORPI, M., & PRESCOTT, B. (1984, April 23–27). *Testing in the secondary schools: Student perspectives.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

HERMAN, J. L., & DORR-BREMME, D. W. (1983). Uses of testing in the schools: A national profile. In W. E. Hathaway (Ed.), *New directions for testing and measurement. No. 19, Testing in the schools* (pp. 7–17). San Francisco: Jossey-Bass.

HINCKLEY, R. H. (Ed.), BEAL, R. S., BREGLIO, V. J., HAERTEL, E. H., & WILEY, D. E. (1979, January). *Report No. 4: Student home environment, educational achievement, and compensatory education* (Tech. Rep. No. 4 from the Study of the Sustaining Effects of Compensatory Education on Basic Skills). Santa Ana, CA: Decima Research.

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

MCCLELLAND, D. C. (1973). Testing for competence rather than "intelligence." *American Psychologist, 28,* 1–14.

MESSICK, S. M. (1980). *The effectiveness of coaching for the SAT: Review and reanalysis of research from the fifties to the FTC.* Princeton, NJ: Educational Testing Service.

MESSICK, S. M. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21,* 215–237.

MILLMAN, J. (1981). Student achievement as a measure of teacher competence. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 146–166). Beverly Hills, CA: Sage.

MITCHELL, D. E., & ENCARNATION, D. J. (1984). Alternative state policy mechanisms for influencing school performance. *Educational Researcher, 13*(5), 1–14.

NATRIELLO, G., & DORNBUSCH, S. M. (1984). *Teacher evaluative standards and student effort.* New York: Longman.

QUINTO, F., & MCKENNA, B. (1977). *Alternatives to standardized testing.* Washington, DC: National Education Association.

STIGGINS, R. J., & BRIDGEFORD, N. J. (1984). *The use of performance assessment in the classroom.* Portland, OR: Northwest Regional Educational Laboratory.

WELLISCH, J. B., MACQUEEN, A. H., CARRIERE, R. A., & DUCK, F. A. (1978). School management and organization in effective schools. *Sociology of Education, 51,* 211–226.

WILEY, D. E., & HARNISCHFEGER, A. (1974). Explosion of a myth: Quantity of schooling and exposure to instruction, major educational vehicles. *Educational Researcher, 4*(3), 7–11.

WRIGHT, B. D., & STONE, M. H. (1979). *Best test design.* Chicago: Mesa Press.

Author

Edward Haertel, Professor, School of Education, Stanford University, Stanford, CA 94305. *Specializations:* Educational measurement, evaluation, psychometrics.