1990

8. Limitations of Using Student-Achievement Data for Career-Ladder Promotions and Merit-Pay Decisions

Ronald A. Berk
The Johns Hopkins University, rberk1@jhu.edu

8

Limitations of Using Student-Achievement Data for Career-Ladder Promotions and Merit-Pay Decisions

Ronald A. Berk
The Johns Hopkins University

INTRODUCTION

A study of U.S. school districts conducted 70 years ago reported that 48% of the districts sampled used merit pay (Evendon, 1918). Since then, the quantity as well as quality of teacher-compensation systems has fluctuated markedly (for details, see Cohen & Murnane, 1985; Murnane & Cohen, 1986; Porwoll, 1979). At present, 29 states are implementing large-scale teacher-incentive programs (a.k.a. career ladder, merit pay, pay for performance), funding local plans, piloting testing models, or using state board of education or legislative mandates to develop programs for teachers and administrators (Southern Regional Education Board, 1986) The status of these programs is summarized in Table 8.1.

Teacher performance is at the core of all of the programs in operation or those being considered. Determining who will receive the pay bonuses, which typically range from $1,000 to $3,000 per year, or be promoted up the career-ladder hinges on the methods used to evaluate teacher performance. The current trend in measurement procedures is to deemphasize supervisory ratings by the building principal and instead to emphasize peer evaluation,

TABLE 8.1
Survey of Teacher Incentive Programs

State	*Local Initiative Only*	*Pilots with State Funding and/or Assistance*	*Full Implementation of State Program*	*State Program Under Development*	*Discussion No Legislative Action Pending*	*Type of Program*
Alabama				X		Career ladder
Alaska						
Arizona		X				Career ladder
Arkansas		(Not Funded)				Career development
California			X			Mentor teacher
Colorado		X				Teacher incentive/career ladder
Connecticut		X				Teacher incentive
Delaware		X				Career development
Florida			X(1)	X(2)		(1) School incentive; (2) Career ladder
Georgia				X		Career ladder
Hawaii					X	
Idaho				(Not funded)		Career compensation
Illinois		X				Teacher incentive
Indiana		X				Teacher incentive
Iowa						
Kansas	X					Teacher incentive
Kentucky		X				Career ladder
Louisiana		X				Career ladder/school incentive
Maine		X				Tiered certification incentive
Maryland	X					Career development incentive
Massachusetts			X			Teacher incentive
Michigan					X	
Minnesota	X					Teacher incentive
Mississippi			X			Teacher incentive

Missouri			X			Career ladder
Montana						
Nebraska				X		Career ladder
Nevada					X	
New Hampshire	X					Teacher incentive
New Jersey			X			Teacher incentive
New Mexico					X	
New York			X			Teacher incentive
North Carolina		X				Career ladder
North Dakota						
Ohio					X	Career ladder
Oklahoma	X					Teacher incentive
Oregon	X					Teacher incentive
Pennsylvania			X			Teacher incentive
Rhode Island	X					Teacher incentive
South Carolina		X(1)	X(2)			(1) Teacher incentive; (2) School incentive
South Dakota					X	
Tennessee			X			Career ladder
Texas			X			Career ladder
Utah			X			Career ladder
Vermont	X					Teacher incentive
Virginia		X				Career ladder/teacher incentive
Washington			X			Mentor teacher
West Virginia				X		Teacher incentive
Wisconsin		X				Career ladder/teacher incentive
Wyoming					X	

Note. Reprinted with permission of the Southern Regional Education Board (1986, p. 9).

classroom observation, student-achievement outcomes, and questionnaire data from principals, teachers, and students (for details, see Southern Regional Education Board, 1986).

Use of Student-Achievement Data

One particular procedure that seems to be gaining acceptance increasingly by legislators and the professionals who are designing the programs is the use of student-achievement data (cf. Robinson, 1983; 1984). These data provide information different from the other measurement tools previously noted. Where classroom observation and ratings by principals, teachers, and students measure a teacher's behavior on the job, student achievement relates to the outcomes of that behavior. That is, the former methods are direct measures of teacher performance; the latter is an indirect measure. Student outcomes are perceived as evidence of a teacher's effectiveness. Because superior teacher performance is the criterion in teacher-incentive programs, the psychometric issue becomes how best to measure that performance—use direct measures, indirect measures, or a combination of both.

Teacher-incentive programs that rely on student-achievement gains have been referred to as "new style merit pay" (Bacharach, Lipsky, & Shedd, 1984), as opposed to "old style merit pay," which bases teacher pay bonuses on principals' evaluations. In 1983, a national survey of merit-pay programs reported that nine school districts in seven states (Arizona, New Hampshire, North Carolina, Oklahoma, South Dakota, Texas, Utah) used student-test scores as evaluative criteria in determining merit pay for classroom teachers (Calhoun & Protheroe, 1983). In all but two of the districts (Dallas and Houston) student achievement served as the only evidence of teacher performance. Today student achievement is a criterion of teacher performance in one third of all statewide teacher incentive/school incentive/career ladder programs. Those programs have been fully implemented in four states (Florida, South Carolina, Tennessee, Utah), are at the pilot stage in four states (Arizona, Kentucky, Maine, South Carolina), and are under development in three states (Alabama, Florida, Georgia). A school incentive program based on student achievement is also under consideration in Alaska, and several career-ladder or merit-pay programs based on student performance have been implemented by local districts (e.g., Campbell County and Danville, Virginia).

Although the results of these surveys do not indicate that the use of student-achievement data is a dominant characteristic or even a trend in most teacher-incentive programs (cf. Moore, 1984), those states where student performance is stressed as the indicator of superior teaching should seriously reconsider the choice of that criterion. Such teacher incentive programs require that students have to perform well on selected achievement tests in order for their teacher to be promoted and/or receive a pay bonus. The teacher's performance on the job may or may not be measured directly. If it is measured, the data are not weighed as heavily in the promotion decision because they are considered "subjective," as compared to the students' achievement data, which are regarded as "objective" evidence of a teacher's performance and effectiveness.

In a more serious application of student-outcome data, student-achievement gains have been used as a major criterion for evaluating teachers as "satisfactory" or "unsatisfactory" in St. Louis. An unsatisfactory classification results in probationary status and can lead to termination. A class action suit was filed in 1986 by the St. Louis Teachers Union (AFT) against this method of teacher evaluation. A U.S. district court decision has not yet been rendered.

Computation of Achievement Gain

When student achievement is adopted as a criterion of teacher performance, it may be expressed as a level of "expected achievement" at the school level to provide school-based awards (e.g., Florida), or as an average pretest-posttest gain score. The last approach, which is most frequently employed in the teacher incentive/career-ladder programs cited previously, is perceived as the simplest, most efficient, and most cost-effective model. It involves a pretest-posttest design where a student-achievement test is administered twice: once at the beginning of the school year (September or October) and once at the end of the year (May or June). One test form or parallel forms may be used. The differences in student performance between the pretest and posttest are computed, and the resulting mean gain score is used to infer the level of teacher performance. Alternatively, the percentage of students in a class who gained "10 or more months in achievement," as measured in grade-equivalent scores, also serves as an index of teacher performance.

Rewarding superior teacher performance on the basis of student-achievement gains is derived from the notion that such gains represent the most concrete product of effective teaching. Proponents of this approach often compare the measurement of a teacher's performance to that of a factory worker's performance; both can be evaluated according to his or her productivity.

Factory Worker–Teacher Productivity Analogy

What's wrong with basing promotions and pay bonuses for teachers on student-achievement gains? Isn't student gain the most important product or work outcome of the job of teaching? After all, if a factory worker's performance can be measured in terms of productivity by the number of widgets he or she produces over a given period of time, why not evaluate a teacher's performance in terms of effectiveness or productivity by his or her students' achievement gains at the end of the school year (cf. Medley, Coker, & Soar, 1984, p. 33)?

The arguments for this factory worker–teacher productivity analogy are derived from the principles of a piece-rate compensation system (Murnane & Cohen, 1986). Piece-rate contracts, where a worker is paid according to the number of widgets produced, is the most common form of "payment by results" (Pencavel, 1977). About 30% of the workers in U.S. factories are employed under piece-rate contracts (Seiler, 1984). These contracts provide a strong incentive for workers to produce, because high productivity results in immediate rewards.

When this piece-rate compensation system is applied to teachers, it breaks down because of the nature of the teaching process and the classroom environment. First, a factory worker uses the same materials (e.g., plywood and chewing gum) to make each product (e.g., widgets). Teachers must work with students whose individual characteristics vary considerably within a single class. This variability precludes all of the students from achieving the same amount at the same rate over 10 months. Second, the characteristics of a factory worker's materials rarely influence his or her skills and rate of production. The worker's ability to build a widget is not affected by the plywood or chewing gum; the quality and quantity of widget production can be attributed solely to the worker. These properties do not generalize to the teaching-learning process. Certain key characteristics of students, such as intelligence and home environment, markedly influence the quality and quantity of their academic achievement, irrespective of what the teacher does in the classroom. Consequently, a teacher's effectiveness is directly affected by the characteristics of the class, which are beyond a teacher's control.

Objectivity of Student-Achievement Data

Students' achievement-test–score gains are often preferred to administrators' ratings of performance and classroom observations because the measurement is perceived to be more objective. This objectivity, however, is illusory. Although students' responses to multiple-choice test items can be scored objectively, the inferences drawn from their scores are subjective. All scores are interpreted, and judgments about student performance are inescapable. When the students' scores are used to infer their teacher's performance, that inference can be erroneous, inasmuch as student achievement is not attributable solely to the teacher. Numerous factors affect the students' performance, only one of which is the teacher's performance.

Assessing superior teacher performance in order to make promotion decisions and award pay bonuses requires a plan that is fair and equitable to all teachers. Establishing such a plan on the basis of achievement-test gains is fraught with difficulty. The difficulties stem primarily from limitations in the testing technology, from factors that influence a teacher's effectiveness beyond his or her control, and from the unfeasibility of executing rigorous experimental-design procedures in the natural school setting (see Haertel, 1986).

This chapter identifies the major limitations of using student achievement as a criterion of teacher performance. It is organized according to four topics: (a) professional and legal standards, (b) factors that influence a teacher's effectiveness beyond his or her control, (c) analysis of achievement gain, and (d) criterion for superior teacher performance.

PROFESSIONAL AND LEGAL STANDARDS

Are there any standards that professionals can use to guide measurement practices in teacher incentive programs? Yes, there are four sources that should be consulted on this question: (a) *Standards for Educational and Psychological Testing* (American Educational Research Association et al., [AERA, APA, NCME], 1985); (b) *Personnel Evaluation Standards* (Joint Committee on Standards for

Educational Evaluation, 1988); (c) *Uniform Guidelines on Employee Selection Procedures* (U.S. Equal Employment Opportunity Commission et al., 1978); and (d) court cases that have relied on the *Guidelines* for the decisions rendered. Although these sources furnish detailed criteria on what should be done, this section concentrates on whether there are any standards that address the use of student-achievement data in the context of teacher evaluation. In addition, it will attempt to extract from those sources the most professionally and legally defensible strategy to evaluate teacher performance.

Standards for Educational and Psychological Testing

Among the four sources, the first set of standards contains one standard that directly attacks the issue. Standard 12.7 states:

> Evaluations of service providers (e.g., teachers and health and social service staff) and administrators should not rest exclusively on the test scores of those people that they serve. (*Primary*)
>
> *Comment:*
>
> Test scores of individuals served (e.g., students) will be affected by a great many factors not directly related to the quality of service they receive. (AERA, APA, NCME, 1985, p. 69)

This standard stipulates that student test scores should not be used as the only criterion to evaluate teachers due to numerous uncontrolled factors that do not relate to teacher performance. (These factors are described in detail in subsequent sections of the chapter.)

Because standardized norm-referenced tests as well as criterion-referenced tests are being considered as the measures of "teacher performance," Standard 6.3, which relates to the validity of test score use, is pertinent:

> When a test is to be used for a purpose for which it has not been previously validated, or for which there is no supported claim for validity, the user is responsible for providing evidence of validity. (*Primary*)

> *Comment:*
>
> The individual who makes the claim for validity is responsible for providing the necessary evidence. Evidence of validity sufficient for test use may often be obtained from a well-documented manual. If previous evidence is not sufficient, then additional data should be collected. (AERA, APA, NCME, 1985, p. 42)

This standard raises the issue of using a student-achievement test to measure teacher performance. An inference about teacher performance is being drawn from the scores on a test designed to measure student achievement. In the test manuals of the major standardized achievement-test batteries published by CTB/McGraw Hill, The Psychological Corporation, Riverside Publishing, and Science Research Associates, not only is no validity evidence provided for using the scores to infer teacher performance, but there is no mention of any intent that the results of the test should be used to evaluate teachers. Consequently, according to Standard 6.3, the burden for gathering appropriate validity evidence rests with the user—the state or local district. The states and districts identified previously have made no visible effort to obtain that evidence.

Other standards germane to the topic of teacher-performance evaluation fall under the sections entitled "Employment Testing" and "Professional and Occupational Licensure and Certification." The technical procedures for evaluating teachers for career-ladder promotion decisions or for retention, demotion, or termination decisions are derived from the same foundation—a comprehensive job analysis that describes the knowledge, skills, abilities, or other personal characteristics necessary to perform the job. The level of performance desired (e.g., average or superior) or expected (e.g., minimum) can be designated in this definition of the job-content domain. The importance of this first step in establishing the content validity of a test that measures teacher performance is expressed in Standards 10.4, 10.5, and 10.6 (AERA, APA, NCME, 1985):

> *Standard 10.4*
>
> Content validation should be based on a thorough and explicit definition of the content domain of interest. For job selection, classification, and promotion, the characterization of the domain should be based on a job analysis. (*Conditional*) (p. 60)

> *Standard 10.5*
>
> When the content-related validation evidence is to stand as support for the use of a test in selection or promotion, a close link between test content and job content should be demonstrated. (*Primary*) (p. 61)
>
> *Standard 10.6*
>
> When content-related evidence of validity is presented, the rationale for defining and describing a specific job content domain in a particular way (e.g., in terms of tasks to be performed or knowledge, skills, abilities, or other personal characteristics) should be stated clearly. The rationale should establish that the knowledge, skills, and abilities said to define the domain are the major determinants of proficiency in that domain. (*Primary*)
>
> *Comment:*
>
> When content-related evidence of validity is presented for a job or class of jobs, the evidence should include a description of the major job characteristics that a test is meant to sample, including the relative frequency and criticality of the elements. (p. 61)

These standards state clearly that a test that measures job performance should be derived from a job analysis and that a close link should exist between the content of the test and the content of the job.

How then would an achievement test of student performance satisfy these standards as a measure of teacher performance? It would be inadequate, because the *Standards* require that job performance be measured *directly* by a test of job content. Students' achievement cannot be used to measure the knowledge, skills, and abilities of a teacher; it does not directly assess a teacher's performance on the job.

Personnel Evaluation Standards

These standards focus exclusively on *personnel evaluation,* defined as "the systematic assessment of a person's performance and/or qualifications in relation to a professional role and some specified and defensible institutional purpose" (Joint Committee on Standards for Educational Evaluation, 1988, pp. 7–8). A standard is "a principle commonly agreed to by people engaged in the professional practice of evaluation for the measurement of the value or the quality of an evaluation" (Joint Committee on Standards for Educational Evaluation, 1981, p. 12). In other words, the *Standards* is the product of a broad search for consensus on what is good and desirable in the evaluation of educational personnel.

Interestingly, among the 21 standards and guidelines for conducting evaluations of teachers, counselors, administrators, and other professional personnel which appear in this document, there is no mention of student-achievement tests. The approach to evaluation advanced in these *Standards* is consistent with the strategy required in Standards 10.4, 10.5, and 10.6, described previously.

The job analysis is the first step. Standard A1 on "Defined Role," with its rationale and guidelines, lays the foundation for the measurement process (Joint Committee on Standards for Education Evaluation, 1988):

> *Standard*
> The role, responsibilities, performance objectives, and needed qualifications of the evaluatee should be clearly defined, so that the evaluator can determine valid assessment criteria. (p. 85)
>
> *Rationale*
> This standard specifies the crucial foundation step in any personnel evaluation process. A carefully developed and sufficiently detailed and delineated description of the role, responsibilities, performance objectives, and qualifications is prerequisite to specifying relevant assessment criteria. (p. 86)
>
> *Guidelines*
> A. Develop job descriptions based on systematic job analysis.
> B. Obtain position description information from as many knowledgeable sources as possible.
> C. Define duties that reflect the needs of students, constituency, and the employing institution.
> D. Specify in detail significant role behaviors, tasks, duties, responsibilities, and performance objectives.
> E. Make clear the relative importance and performance level of each standard used to define success in the position.
> F. Investigate and resolve any discrepancies in the position description.
> G. Make clear the relationship between performance indicators and the standard with which each indicator is associated. (pp. 86–87)

The teaching environment and the factors that can influence or constrain teacher performance are considered in Standard A2 on "Work Environment":

> *Standard*
> The context in which the evaluatee works should be identified, described, and recorded so that environmental influences and constraints on performance can be considered in the evaluation. (p. 90)
>
> *Rationale*
> Holding educators accountable for the effects of variables they cannot control or influence is likely to lead to resentment and low morale. Failure to take account of environmental factors may also threaten the validity of the evaluation process. (p. 90)
>
> *Guidelines*
> A. Identify and record contextual variables that might affect the work environment.
> B. Consider available resources, working conditions, community expectations, and other context variables that might have affected performance. (p. 91)

The validity issue in personnel evaluation is given attention in Standard A4 on "Valid Measurement":

> *Standard*
> The measurement procedures should be chosen or developed and implemented on the basis of the described role and the intended use, so that the inferences concerning the evaluatee are valid and accurate. (p. 98)
>
> *Rationale*
> Validity is the single most important issue in the assessment of any evaluation process. If the evaluation is to serve its intended purpose, then the inferences and judgments that are made must be defensible. The selection, development, and implementation of the instruments and procedures for collecting information, as well as the basis for synthesizing the information and drawing inferences from it, must be clearly linked to the purposes for which judgments, inferences, and decisions are made. Further, these linkages must be documented and made public. (p. 99)

One of the common errors listed in relation to the guidelines for Standard A4 is "using a measurement procedure for multiple purposes when it is valid for only one, for example, using students' scores on a nationally standardized test to assess the performance of a teacher or administrator when the test has not been validated for the latter purpose" (Joint Committee on Standards for Educational Evaluation, 1988, p. 100)

Reliability is assigned similar weight in Standard A5 on "Reliable Measurement":

> Measurement procedures should be chosen or developed and implemented to assure reliability, so that the information obtained will provide consistent indications of the performance of the evaluatee. (p. 104)

The preceding standards plus many others in the document stress appropriate, technically defensible, and professionally acceptable practices for evaluating teacher performance. These up-to-date standards do not recommend the applicability of student test scores in this context.

Uniform Guidelines on Employee-Selection Procedures

In addition to the sets of professional standards cited in the first two sections, there are government regulations that protect individuals against any form of employment discrimination. Title VII of the 1964 Equal Employment Opportunity Act is enforced by the Equal Employment Opportunity Commission (EEOC) based on a set of guidelines, entitled *Uniform Guidelines on Employee Selection Procedures* (U.S. Equal Employment Opportunity Commission et al., 1978). These *Guidelines* apply to every kind of personnel-assessment technique used to make an employment decision. This includes "any measure, combination of measures, or procedures used as a basis for any employment decision" (p. 38308).

The purpose of the *Guidelines* is described in Section 1B:

> These guidelines incorporate a single set of principles which are designed to assist employers, labor organizations, employment agencies, and licensing and certification boards to comply with requirements of Federal law prohibiting employment practices which discriminate on grounds of race, color, religion, sex and national origin. They are designed to provide a framework for determining the proper use of tests and other selection procedures. (p. 38296)

One primary concern of the EEOC is whether an assessment procedure results in adverse impact against members of a racial, ethnic, or sex group. The EEOC would consider that a test that has no adverse impact complies with Title VII. If adverse impact is found, it would have to be justified in terms of appropriate validity evidence.

Suppose a disproportionate number of Black teachers in a local district were denied career-ladder promotions or were placed on probation because their evaluations were unsatisfactory compared to those of the White teachers. The determination of adverse impact and compliance with the *Guidelines* by the EEOC would hinge on the validity evidence that supports the use of the particular measurement tools for those "employment decisions."

What types of validity evidence must be documented? The *Guidelines* indicate the same types of evidence as those needed to satisfy the validity standards cited previously, where the most crucial step is the job analysis. The *Guidelines* (U.S. Equal Employment Commission et al., 1978) specify validity studies for (a) content validity—"an analysis of the important work behavior(s) required for successful performance and their relative importance and, if the behavior results in work product(s), an analysis of the work product(s)" (sec. 14C [2]); (b) construct validity—"the job analysis should show the work behavior(s) required for successful performance of the job, . . . the critical or important work behavior(s) in the job or group of jobs being studied, and an identification of the construct(s) believed to underlie successful performance of these critical or important work behaviors in the job or jobs in question" (sec. 14D [2]); and (c) criterion-related validity—"to determine measures of work behavior(s) or performance that are relevant to the job or group of jobs in question" (sec. 14B [2]).

Because student-achievement gain is perceived as an outcome of teaching, that is, work outcome, why not use achievement as a criterion variable? The *Guidelines'* definition of criteria for criterion-related validity studies is as follows:

> Whatever criteria are used should represent important or critical work behaviors(s) or work outcomes. Certain criteria may be used without a full job analysis if the user can show the importance of the criteria to the particular employment context. These criteria include but are not limited to production rate, error rate, tardiness, absenteeism, and length of service. A standardized rating of overall work performance may be used where a study of the job shows that it is an appropriate criterion. (pp. 38300–38301)

Notice that all except one of the preceding criteria stated are objective, single measures of the person being evaluated. Achievement gain, however, is a collective (class) index representing the performance of individuals with diverse academic (and usually demographic) characteristics, which is then applied to the teacher being evaluated. Despite the common interpretation of student-achievement gain as the direct product or outcome of teaching, as noted in the previous section, gain is an indirect measure of teacher performance.

Court Cases

The court cases that have implications for teacher evaluation and for the use of student achievement data to assess teacher performance can be classified into general employment decisions and teacher employment decisions. The purpose of this section is to extract from the court decisions the key factors or issues that are germane to the student test-score approach to teacher evaluation.

General-Employment Decisions

There are numerous court cases involving the use of tests and other measurement techniques in a variety of employment applications that may have a bearing on future litigation on teacher evaluation (see Madaus, chap. 7). Excellent reviews of these cases have been completed by Bernardin and Cascio (1984) and Nathan and Cascio (1986). Their reviews suggest that the courts have been guided by a number of factors in assessing personnel-evaluation systems; some relate to technical standards such as those stated in the *Guidelines,* whereas others pertain to proper personnel practices that help to safeguard against discriminatory employment decisions (Nathan & Cascio, 1986). Four particular factors have emerged from the reviews of Cascio and Bernardin (1981) and Bernardin and Beatty (1984):

1. Standards for performance should be based on a *job analysis.*
2. Evaluation should be based on *specific job dimensions,* not on a global or overall measure.
3. Ratings should be made on behaviorally based *performance dimensions* rather than on personality traits.

4. *Documentation* should be kept and should be accurate. Kleiman and Durham (1981) also emphasized the evidence essential to demonstrate that a performance evaluation is valid or job related. Further, they recommend presenting evidence that the evaluation procedures do not discriminate.

Consistent with the *Standards* and the *Guidelines,* the courts have affirmed the importance of a thorough job analysis. In the teacher-evaluation literature, empirically based schemes have been developed to identify specific job dimensions and behaviorally based performance dimensions (see review by Medley et al., 1984, chap. 4). The methods for assessing these dimensions, however, should be direct rather than indirect. The courts have supported the use of ratings of behavior as the basis for performance evaluation. There is no precedent for the use of student-test scores to measure teacher performance.

Teacher-Employment Decisions

Strike and Bull (1981) surveyed federal law and state regulations governing teacher evaluation, especially personnel policies and actions that relate to termination, salary determination, and promotion and demotion. They recommended that teacher-evaluation procedures focus "only on those aspects of a teacher's performance, behavior, and activities that are directly or indirectly relevant to the teacher's ability to execute the legitimate responsibilities that attach to the job" (p. 336). Their conclusions regarding the principle of *evaluative relevance,* however, are most appropriate to the issues of interest:

> The relevance requirement for . . . external information is . . . connected with the legal core of meaning of teaching competence: external information must be plausibly indicative of the teacher's capacity to fulfill central instructional responsibilities. . . . [C]ertain indirect measures of teaching ability, such as student test results, teacher tests, or research-based instruments, may be held legally relevant to judgments of competence under a variety of conditions. (p. 337)

This principle indicates that student-test scores may be legally relevant to the evaluation of teaching competence. The most recent test of evaluative relevance is *St. Louis Teachers Union v. Board of Education of St. Louis,* described previously, for which a decision has not yet been rendered.

Summary

The themes that recur in both sets of *Standards,* the *Guidelines,* and the court cases are as follows:

1. A comprehensive job analysis is crucial.
2. Evidence of job relatedness for all evaluation instruments must be provided.
3. Appropriate evidence of validity and reliability of test or scale scores used for employment decisions must be obtained.
4. Evidence that instruments are unbiased and nondiscriminatory of racial, sex, and ethnic subpopulations should be available.

As these themes are applied to teacher-incentive programs, it is clear that teacher performance should be measured directly in terms of on-the-job behaviors. An indirect measure such as student-achievement performance may be legally relevant and appropriate as one among several evaluative criteria, although not defensible according to the *Personnel Evaluation Standards* (Joint Committee on Standards for Educational Evaluation, 1988, p. 100).

FACTORS THAT INFLUENCE A TEACHER'S EFFECTIVENESS BEYOND HIS OR HER CONTROL

In the preceding section, it was noted that one of the intractable problems of using student achievement to measure teacher performance is isolating teacher performance as the primary explanation for changes in student performance. This issue is addressed specifically by Standard 12.7 (AERA, APA, NCME, 1985, p. 69) and Standard A2 (Joint Committee on Standards for Educational Evaluation, 1988, p. 114).

There are several factors that can influence a teacher's measured effectiveness that are beyond his or her control. These factors can account for a sizable proportion of the gain that may be exhibited in student achievement. Many of the factors have been identified previously by Berk (1984c; 1988), Haertel (1986), and Medley et al. (1984). In addition, several reviews of research on input-output analyses of schools by Bridge, Judd, and Moock (1979), Centra and Potter (1980), Cohn and Millman (1975), and, especially, Glasman and Biniaminov (1981) provide valuable insights into the impact of numerous variables on achievement. The last review is the most comprehensive to date.

The factors examined in this corpus of literature cluster into three categories: (a) student characteristics, (b) school characteristics, and (c) test characteristics. The work of Glasman and Biniaminov (1981) addresses most of the characteristics that fall into categories a and b; the issues that relate to category c have been discussed by Berk (1988).

Student Characteristics

There are at least seven types of student characteristics that can positively or negatively affect student achievement: (a) intelligence, (b) attitude, (c) socioeconomic level, (d) race/ethnicity, (d) sex, (e) age, and (f) attendance. These are attribute variables. Students possess these characteristics when they enter the classroom; most of them cannot be manipulated by the teacher. Under experimental conditions it might be possible to change intelligence and attitudes to some degree, or to improve attendance. However, under normal nonexperimental conditions, a teacher is assigned a class of students with a given set of characteristics.

The aforementioned student characteristics are described briefly in this section to determine the degree and direction of their effect on student achievement.

Intelligence. Intelligence or academic aptitude typically correlates from 0.40 to 0.70 with achievement, as measured by standardized test batteries. When the correlations are based on class means, they may be as high as 0.90 (Soar & Soar, 1975). As Medley et al. (1984) pointed out, "a correlation of .90 means . . . that about 80 percent of the differences in the pupil achievement scores used to evaluate a teacher were present before [he or] she had any chance to influence them" (p. 34). Furthermore, the interaction of intelligence with other student characteristics and school characteristics can affect achievement levels (Cronbach & Snow, 1977).

Attitude (three variables). Three types of student attitude have been investigated: (a) locus of control—the extent to which outcomes are attributed to self-action (internals) or to fate, chance, and powerful others (externals), (b) self-concept—the beliefs about one's personal characteristics, and (c) academic aspiration—the motivation to achieve in school. Glasman and Biniaminov's (1981) synthesis of the research indicates consistent findings that internal control, high self-concept, and high academic aspirations positively influence reading and mathematics achievement. Locus of control and self-concept tend to have a much stronger effect on achievement than academic aspirations, and these attitudes are stronger determinants of verbal achievement than of socioeconomic variables (Mayeske & Beaton, 1975).

Socioeconomic level (six variables). Six family-background variables have been used in combination to define socioeconomic level, including family size, family income, family occupational status, family possessions, parental education, and family's educational environment. The results of 17 studies were consistent: All of these components of socioeconomic level except family size were strongly and positively correlated with reading, mathematics, verbal, and composite achievement (see Glasman & Biniaminov, 1981). Family size was negatively correlated with achievement (e.g., Hanushek, 1972; Wiley, 1976).

Race/ethnicity. Racial composition of elementary and secondary schools defined either as percentages of White, Black, or non-White students or as a dummy coded variable (Black = 1, others = 0) was negatively correlated with reading, mathematics, and verbal achievement where there was a majority of Black or non-White students. Only one study by Winkler (1975) found a positive association. Interestingly, Mayeske et al. (1973) reported that race/ethnicity accounted for 24% of the variance in achievement when socioeconomic factors were uncontrolled, and only 1% when those factors were controlled.

Sex. Several studies of the relationship between sex, coded as female = 1 and male = 0, and achievement have found consistently positive correlations with reading and composite achievement and negative correlations with mathematics (e.g., Michaelson, 1970; Summers & Wolfe, 1977). In other words, females perform better in reading and males better in math.

Age. Three studies that examined the variable of age in grade, coded as over-age = 1 and not over-age = 0, at the elementary and secondary levels reported negative correlations with reading and mathematics achievement (Boardman, Davis, & Sanday, 1974; Levin, 1970; Michaelson, 1970). Consequently, the age composition of a class can affect achievement gains negatively to the extent that over-age students are in the majority.

Attendance (four variables). Student attendance has been expressed as student turnover, days present, quantity of schooling index, and student unexcused absences and lateness. Only three studies have explored this issue. Their findings at the elementary level indicate that poor attendance negatively affects reading, mathematics, and composite achievement (Murnane, 1975; Summers & Wolfe, 1977; Wiley, 1976).

School Characteristics

Beyond the characteristics of students which can affect achievement gains, there are numerous variables of school conditions and instructional personnel that exhibit similar effects. These variables have been analyzed by Glasman and Biniaminov (1981) and Haertel (1986).

School Conditions. More than 25 studies have investigated variables that relate to school services, facilities, expenditures, staff, and climate. They include the following:

1. school library (number of books per student)
2. class size (number of students per classroom)
3. size of a type of class (e.g., mean school class size in math)
4. age of building
5. size of school site
6. size of school enrollment
7. size of staff
8. turnover of staff
9. expenditures
10. quality of instructional materials and equipment (e.g., desks, chalkboards, textbooks, computers)
11. schoolwide learning climate
12. instructional support (e.g., aides, resource teachers, team teaching)

Glasman and Biniaminov's (1981) review of research on variables 1 through 8 led to their conclusion that the direction and significance of those variables' effects on achievement were inconsistent; the results were positive, negative, and mixed. However, there were consistent negative correlations between class and school size and reading and mathematics achievement; school library size was also positively associated with reading achievement.

Expenditures (variable 9) for administration, instruction, and extracurricular activities were positively correlated with reading and composite achievement (Benson et al., 1965; Cohn & Millman, 1975; Kiesling, 1969; 1970). Research on variables 10 through 12 was examined by Haertel (1986). He concluded that (a) quality of instructional materials may influence achievement (Wiley & Harnischfeger, 1974); (b) teachers can be more effective in schools with favorable learning climates (Bridge et al., 1979; Brookover, Beady, Flood, Schweitzer, & Wisenbaker, 1979); and (c) instructional support at the elementary and secondary levels can affect student performance.

Instructional Personnel. There are several teacher-background and personal characteristics and teacher-assignment and attitude variables that influence student achievement. These variables include:

1. education degree
2. undergraduate education type
3. teaching experience
4. verbal achievement
5. race
6. sex
7. teaching load
8. time in discipline
9. job satisfaction

In their review of more than 20 studies of these variables Glasman and Biniaminov (1981) concluded: (a) higher levels of education, verbal achievement, and experience affected reading and mathematics achievement positively, (b) increased teaching loads and time in discipline produced negative effects on reading, mathematics, and verbal achievement, and (c) greater job satisfaction was positively correlated with reading, mathematics, and verbal achievement.

Test Characteristics

Although the 17 student characteristics and 21 school characteristics identified thus far should suggest the difficulty of attributing student-achievement gains to teacher performance, just how that achievement is measured is equally important to the teacher-evaluation process. The characteristics of the achievement test selected can have a profound effect on what is actually measured, how it is interpreted, and the extent to which student performance reflects teacher effectiveness. In this section, pertinent test characteristics are described under three topics: (a) type of achievement test, (b) curricular and instructional validity, and (c) test score metric.

Type of Achievement Test. The first decision that must be made is the type of achievement test(s) to be used to measure teacher performance. The choices often reduce to standardized norm-referenced tests and criterion-referenced tests. The selection of any single test should be based on its technical adequacy in terms of norms, validity, and reliability. Standards and criteria for judging adequacy are set forth in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1985). Special attention should be given to the characteristics of curricular and instructional validity. It is important that the items on the test match the objectives of the local curriculum and the instruction that actually occurs. Tests that are insensitive to what is taught in any subject area are inappropriate measures of student achievement as well as teacher performance.

Because standardized norm-referenced tests, such as the Iowa Tests of Basic Skills, California Achievement Tests, Comprehensive Tests of Basic Skills, Metropolitan Achievement Tests, Stanford Achievement Test, and Survey of Basic Skills, typically survey broad domains of content, they rarely "mirror a particular curriculum." In fact, the tests are expressly designed to minimize local, state, and regional content biases (Green, 1983; Mehrens, 1984). If the achievement-test scores do not accurately measure achievement in the program, their validity is weakened. The degree of invalidity is contingent upon the match between what the test measures and what the curriculum covers.

In contrast to standardized tests, criterion-referenced competency tests are tailored to measure the instructional objectives of a school-based program (Berk, 1984a). Such tests, however, must be developed by the local or state educational agency, or in collaboration with a professional test-development contractor. Unfortunately, the experiences with minimum-competency test construction over the past decade indicate that the products of local efforts are far from technically adequate (Berk, 1986). Commercially developed criterion-referenced tests have also been plagued by technical deficiencies (Hambleton & Eignor, 1978) related to item characteristics, mastery–nonmastery cut-off scores, and decision consistency.

Curriculum and Instructional Validity. Although content, criterion-related, and construct validities are applicable to achievement-test scores in general, there are specific types of validity evidence that must be obtained to consider drawing inferences about teacher performance. Such evidence relates to curricular and instructional validity.

Curricular validity refers to the extent to which the items on the test measure the content of a local curriculum (cf. McClung, 1979, p. 682). Although conceptually similar to content validity (Madaus, 1983; Schmidt, Porter, Schwille, Floden, & Freeman, 1983) and even viewed by some experts as synonymous with content validity (Cureton, 1951; Hopkins & Stanley, 1981, chap. 4; Madaus, Airasian, Hambleton, Consalvo, & Orlandi, 1982), curricular validity is operationally very different. In the case of standardized norm-referenced tests, it does not focus on the content domain the test was designed to measure; it deals with a specific domain to which the test is later applied. The relevance of the test in a specific application is being evaluated. Rarely would perfect congruence between the two domains ever occur (e.g., Bower, 1982; Gramenz, Johnson, & Jones, 1982; Jenkins & Pany, 1978; Madaus et al., 1982; Porter, Schmidt, Floden, & Freeman, 1978).

Evidence of curricular validity is obtained by determining the degree of congruence or match between the test items and the curriculum. This is based on a systematic, judgmental review of the test against the curricular objectives or materials by content experts. These experts may be classroom teachers or curriculum specialists; they are the only professionals in a position to judge curricular validity. The review can vary as a function of the following: (a) single grade versus cumulative grade content, (b) specificity of objectives or content/process matrix, (c) internal versus external determination, and (d) curricular materials versus actual classroom activities (for details, see Schmidt, 1983a; 1983b; Schmidt et al., 1983). What emerges from this process are several estimates of content overlap, including the amount of content in common, the percentage of the local curriculum measured by the test, and the percentage of items on the test not covered by the curriculum. The second estimate in particular can furnish evidence of the curricular validity of the test.

When a standardized test is found to have low curricular validity, alternative testing procedures should be considered. One procedure involves customizing the test by developing supplementary items to fill in the identified measurement gaps. These items would be administered and scored in conjunction with the standardized test. Technical problems arise in evaluating the validity and reliability of the "supplementary test" and in equating its scores to the appropriate national norms. Another procedure is to choose an out-of-grade-level test that provides a better curricular match.

An important issue related to curricular validity is whether achievement tests measure what is actually taught in the schools. Very often it is simply assumed or implied that evidence of curricular validity means that the objectives guided the instruction and the curricular materials were used in the classroom. This does not necessarily follow, as several studies have demonstrated (Hardy, 1984; Leinhardt & Seewald, 1981; Leinhardt, Zigmond, & Cooley, 1981; Poynor, 1978; Schmidt et al., 1983). What is measured by the test is not always the same as what is taught, especially with regard to standardized tests. Hence, a distinction has been made between these different domains to which the test items can be referenced (Schmidt et al., 1983). When the domain is the instruction actually delivered, a "measure of whether schools are providing students with instruction in the knowledge and skills measured by the test" (McClung, 1979, p. 683) is called instructional validity.

Instructional validity refers to the extent to which the items on the test measure the content actually taught to the students. Several techniques have been proposed for assessing the overlap between the test and the instruction. Popham (1983) identified four data-sources for describing whether students have received instruction that would enable them to perform satisfactorily on a test: (a) observations of classroom transactions, (b) analyses of instructional materials, (c) instructor self-reports, and (d) student self-reports. Although he views these sources as methods for determining the adequacy of test preparation (Yalow & Popham, 1983), they can be considered as techniques for gathering evidence of instructional validity. Unfortunately, Popham's (1983) evaluation of those techniques suggests that the process of estimating the percentage of a standardized test that has been covered by teaching has numerous methodological problems related to executing the data-gathering procedures (see Leinhardt, 1983; Schmidt et al., 1983). They stem, in large part, from the variability of instructional content, not only among different classes, but within a single classroom.

The evidence from an instructional validity study can reveal "content taught but not tested" and "content tested but not taught." Both types of evidence have significant implications for inferring teacher effectiveness from student-achievement gains. In the case of the former, if the evidence indicates that there is a considerable amount of content being taught but not covered by the achievement test, then the students' performance gains may only partially reflect the teacher's performance. Instruction on skills at the higher levels of cognition (e.g., application, analysis), which are the levels rarely measured by standardized norm-referenced tests (Soar & Soar, 1983), might not be assessed. In that case, an inference about a teacher's performance from the achievement test scores would need to be qualified in the context of what was not measured by the test.

Conversely, if there is validity evidence that a proportion of the test items measures content that was not taught to the students, then inadequate achievement gains on that test cannot be attributed to the teacher's performance, unless that particular content was supposed to be taught. The most common strategy to address this type of test content–instruction mismatch is for teachers to teach the objectives measured by the test. If teachers are to be evaluated according to their students' test performance, then it is highly probable that a sizable portion of the instruction will be driven by the test content. Because most achievement tests tend to measure simpler objectives, as opposed to complex or higher-order objectives, teaching will attempt to maximize student progress on those objectives to produce large achievement gains (Medley et al., 1984, chap. 3).

Test-Score Metric

In order to perform basic arithmetic calculations, such as computing the difference between pretest and posttest scores and group-average scores, equal-interval scales are essential. The most frequently used derived-score scale for norm-referenced tests is the *grade equivalent*. It is not an interval scale and has several other serious deficiencies (see Angoff, 1971; Berk, 1984b; Flanagan, 1951;

Horst, 1976; Horst, Tallmadge, & Wood, 1974; Linn, 1981; Williams, 1980). Those deficiencies have been summarized by Berk (1984b):

> Grade equivalents
>
> 1. invite seemingly simple but misleading interpretations;
> 2. assume that the rate of learning is constant throughout the school year;
> 3. yield different growth rates at different score levels;
> 4. are derived primarily from interpolation and extrapolation rather than from real data;
> 5. are virtually meaningless in the upper grade levels for subjects that are not taught at those levels;
> 6. exaggerate the significance of small differences in performance;
> 7. are affected by changes in educational customs regarding promotion from grade to grade;
> 8. vary markedly from publisher to publisher, from test to test, from subtest to subtest within the same test battery, from grade to grade, and from percentile to percentile. (pp. 94–96)

Consistent with these deficiencies are the cautions cited for interpreting grade equivalents in relation to Standard 4.1 of the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1985):

> Test publishers and users can reduce misinterpretations of grade-equivalent scores, for example, by ensuring that such scores are (a) reported only for grades in which actual growth can be expected, (b) reported only for grades for which test data are available, and (c) accompanied by instructions that make clear that grade-equivalent scores do not represent a standard of growth per year or grade and that 50% of the students tested in the standardization sample should by definition fall below grade level, that if a student scores above grade level it does not necessarily mean that the student has mastered the content material of the higher grade level, and that interpretations of differences between grade equivalent scores on separate subtests should be avoided. (p. 33)

Because grade equivalents can distort a student's actual achievement levels on both the pretest and posttest, there is no technically sound reason to justify their use in the estimation of gain scores. As Angoff (1971) noted, "their simplicity is far more apparent than real" (p. 525); however, the adverse consequences of their continued use will be far more real than apparent.

Percentile ranks are also unacceptable for gain-score analysis inasmuch as they comprise an ordinal scale. Although their interpretation is direct and readily understood, the inequality of percentile units on different parts of the scale render them inappropriate for computing pretest-posttest gains.

The preferred metric for gain-score analysis is simple *raw scores*. They are appropriate when the same test form is administered both times. If parallel forms are employed or it is desirable to compare performances from one subtest to another or from class to class, *scaled scores* should be used. These scores possess the property of equal intervals and permit comparisons of tests within and across grade levels.

For criterion-referenced tests, raw score or proportion correct is an appropriate metric to estimate gain. Linn (1981) recommended that if the content domain of the test is explicitly defined and random or stratified random samples of items can be generated, the estimate of proportion correct on each item sample can be used to obtain growth curves.

Summary

This section presented 17 student characteristics, 21 school characteristics, and 4 achievement test characteristics that can influence the evaluation of teacher performance. In other words, there are more than 40 factors that affect student achievement, its measurement, and its interpretation, irrespective of teacher performance. Despite the interrelationships among many of these factors and the efforts to control or eliminate some of them (see Haertel, 1986), an individual teacher whose performance is being measured via achievement gains is rarely in a position to manipulate those factors in order to neutralize their effect on his or her performance. According to the research literature cited previously, most of the factors have a positive effect on achievement and, consequently, could account for a large proportion of the overall gain over 10 months. A few of the factors had negative effects, and other factors could be positive or negative.

ANALYSIS OF ACHIEVEMENT GAIN

In addition to the aforementioned factors that affect student achievement and inferences about teacher performance, the pretest-posttest database for computing gain scores and the inferences drawn from those scores possess other limitations. Typically, the achievement-test database used in some incentive programs focuses on the difference in the students' performance on a standardized achievement test between September (or October) and May (or June) during the same school year; alternatively, the two testings can occur in May of one school year and again in May of the succeeding year. In either case only two measurement points (pretest and posttest) are used.

As noted in the introduction to this chapter, the states and particular school districts who rely on achievement data for promotion and pay bonus decisions compute the difference between the two testings using three methods:

1. Subtract a student's posttest score (X_2) from the pretest score (X_1), or $X_2 - X_1$.

   Calculate the percentage of students who gained (10 months to be "on or above grade level").
2. Average the $X_2 - X_1$ gain scores for a single class (i.e., mean gain score).
3. Average the $X_2 - X_1$ gain scores for an entire grade level in a school.

Methods 1 and 2 are intended to measure teacher performance; method 3 focuses on school effectiveness. A few of the current teacher-incentive programs employ one or any combination of those methods.

This section examines the adequacy of the preceding methods as measures of gain and the validity of inferences from gain scores.

Measurement of Gain

Traditional Deficiencies

During the past 40 years a considerable amount of research has been devoted to the study of how to measure change or gain over time (see Bereiter, 1963; Cronbach & Furby, 1970; Linn & Slinde, 1977; Lord, 1956; 1963; O'Connor, 1972; Webster & Bereiter, 1963). Much of this work has cited two major deficiencies of pretest-posttest gain scores: their low reliability and their negative correlation with pretest scores.

The formula for the *reliability of a gain score* (r_{GS}) can be expressed in terms of the reliabilities of the prescores (r_{11}) and postscores (r_{22}), considered separately, and the correlation between them (r_{12}), or

$$r_{GS} = \frac{r_{11} + r_{22} - 2r_{12}}{2(1 - r_{12})} .$$

Low reliability can result from this formula under certain observable conditions. First, if the alpha reliability coefficients are identical and equal to the test-retest coefficient, the reliability of the gain score is zero. Second, a high test-retest correlation tends to produce a low gain-score reliability. For example, a test with a common variance and a reliability of 0.80 would have a gain score reliability of 0.60, 0.50, 0.33, and 0 when the correlation (r_{12}) was 0.50, 0.60, 0.70, and 0.80, respectively (Linn, 1981, p. 87). Interestingly, these low gain-score reliabilities would rarely occur because the assumption of common variance is not usually upheld in practice.

This low reliability of gain scores has been regarded as a serious concern in individual student decision making and in decisions based on aggregates of individual gain scores (Method 1). The reliability of a mean gain score (Methods 2 and 3) has been viewed as problematic in terms of stability coefficients from one year to the next. From several studies of the stability of class mean gain, it was found that the median stability coefficient was approximately 0.30 (Brophy, 1973; Rosenshine, 1970). This instability of gains occurred across years, teachers, grade levels, subtest-subject areas, and Title I versus non-Title I schools.

The second deficiency of gain scores is their *negative correlation with pretest scores.* This negative bias has been cited as an important reason to avoid gain scores (Linn & Slinde, 1977; O'Connor, 1972). If the pretest- and posttest-score variances are equal, the correlation between the pretest scores and gain scores is necessarily negative because r_{12} will be less than 1.0. This means that students with low pretest scores will tend to have larger gains than students with high pretest scores. However, the converse is possible. If the posttest variance is considerably larger than the pretest variance, r_{12} may be positive, in which case the initially higher scoring students have a built-in advantage (see Linn, 1981; Zimmerman & Williams, 1982).

Deficiencies as Misconceptions

The findings of investigations comparing numerous strategies for estimating gain (e.g., Corder-Bolz, 1978; Overall & Woodward, 1975; 1976; Richards, 1976) and the reanalyses of these issues by Rogosa (1980; Rogosa, Brandt, & Zimowski, 1982; Rogosa & Willett, 1983; 1985) and others (Nesselroade, Stigler, & Baltes, 1980; Willett, 1988; Zimmerman & Williams, 1982) strongly indicate that the aforementioned deficiencies are not serious. Low reliability and negative correlation with initial status are misconceptions rather than deficiencies.

On the problem of low reliability, Rogosa et al. (1982) pointed out: (a) "low reliability [of gain scores] does not necessarily mean lack of precision," and (b) "the difference between two fallible measures can be nearly as reliable as the measures themselves" (p. 744). Overall and Woodward (1975) also demonstrated that the unreliability of gain scores should not be a cause for concern in determining an instructional effect between two testings. A true effect can be evidenced using a *t*-test for paired observations "irrespective of the zero reliability of difference scores upon which *all* calculations are based" (p. 86). In fact, the power of tests of significance is maximum when the reliability of the difference scores is zero.

The negative bias of the correlation should be interpreted as an artifact of measurement error on the estimation of the correlation. Rogosa et al. (1982) argued that the bias is not a fundamental difficulty with the use of the gain score as a measure of change.

Alternative Methods

A variety of methods have been proposed for estimating gain, including raw gain, gain adjusted for pretest error, gain adjusted for pretest and posttest error, the difference between true posttest and pretest scores (Lord, 1956), raw residual gain, estimated true residual gain, a "base-free" procedure (Tucker, Damarin, & Messick, 1966), and posttest score adjusted for initial academic potential. None of these procedures provides a satisfactory solution. Three other approaches supplement the information on the two data points (X_1 and X_2) with between-person information (e.g., reliabilities and measurement error variances): (a) weighted reliability measures, (b) Lord–McNemar regression estimates, and (c) Bayes growth-curve estimates (for details, see Rogosa et al., 1982).

New Directions

Despite all of the research cited in this section, which has addressed the technical problems in measuring gain, the most important deficiency of the pretest-posttest gain score is the meager information it yields based on only two measurement points. This issue was virtually ignored in the research literature until the 1980s. The use of *multiwave data,* where three measurements (September-January-May), four measurements (September-December-March-May), or more are obtained, vastly improves the measurement of change over time simply because additional information on each student is available (Rogosa et al., 1982). Multiple measurements provide greater precision in estimating gain than just two measurements (see Bryk & Raudenbush, 1987; Rogosa & Willett, 1985; Willett, 1988).

Validity of Gain-Score Inferences

The validity of gain-score inferences pertains to the underlying pretest-posttest design. The several possible factors jeopardizing the internal validity of the one-group pretest-posttest design have been discussed extensively in the research methodology literature à la Campbell and Stanley (1966) and Cook and Campbell (1979). They have also been emphasized in reviews of the RMC Research Corporation's Title I evaluation model A (Horst, Tallmadge, & Wood, 1974; Linn, 1979; 1980b; 1981; Linn, Dunbar, Harnisch, & Hastings, 1982; Tallmadge, 1982; Tallmadge & Wood, 1976). Among the factors of history, maturation, testing, instrumentation, statistical regression, selection, mortality, and interactions with selection, only those germane to the inference of teacher performance are described in this section.

The gain score computed from the pretest and posttest administrations is to be attributed to the teacher's performance. The score is one indicant of his or her effectiveness. The validity question asks: What other plausible explanations could account for the gain score? If the gain score is invalidated, such that there are many reasons for the improvement in the students' performance, only one of which may be teacher effort, then promoting a teacher or awarding a pay bonus would be unjustified. The relevance of the alternative explanations for gain may vary across classes, grade levels, subject areas, and schools.

History

Gain may be due to history in the sense that events outside of the school setting could have occurred over the 9 to 10 months between the testings which, in turn, affect student achievement. Home and community resources (e.g., books, computers), which may vary as a function of socioeconomic level, educational and cable television programs, and the like, could influence a student's progress in reading, mathematics, and other subjects, irrespective of what happens in the classroom.

Maturation

As the students grow older, wiser, and more experienced over the school year, their learning and measured achievement will be affected to some degree.

Statistical Regression

Students who have low pretest scores will score higher on the posttest, and students who score high on the pretest will score relatively lower on the posttest. That is, the most extreme scores on the pretest tend to "regress toward the population mean" on the posttest. The regression effect operates (a) to increase obtained pretest-posttest gain scores among low pretest scores, (b) to decrease obtained change scores among students with high pretest scores, and (c) to not affect obtained change scores among scores at the center of the pretest distribution (for details, see Cook & Campbell, 1979, pp. 52–53). These changes that occur due to regression cannot be attributed to the teacher. The magnitude of the changes depends on the test-retest reliability coefficient and the ability distribution in the class at the time of the pretest. The higher the reliability and the more average the students, the less will be the regression effect.

Mortality

In the course of a school year, students can leave a given class for any number of reasons. As the composition of the class changes—some students leave and others transfer in—a selection artifact results. The students taking the posttest may be different from those who took the pretest.

Interactions with Selection

When mean gain scores are compared across classes in one school or across schools to determine which teacher(s) or school(s) deserves a financial award, there are additional factors such as selection-history and selection-maturation that could account for differential gains in the classes or schools. *Selection-history* results when the schools being compared are located in different geographic and socioeconomic areas. The students in each school could experience a different local history that might affect achievement gains. *Selection-maturation* occurs when the students in different classes or schools are maturing at different rates due to differences in socioeconomic background or other variables. As noted previously, socioeconomic level is related to achievement growth rates.

Multiple Sources of Invalidity

Ideally, it would be desirable to partial out of the total gain that proportion of gain attributable to extraneous (noninstructional) factors. Suppose that the observed gain scores by students in a class were expressed in terms of variance components, or

$$\sigma^2_{OG} = \sigma^2_{TG} + \sigma^2_{E};$$

that is, the variance of the observed gain scores (σ^2_{OG}) equals the variance of true gain scores (σ^2_{TG}) plus the variance arising from errors of measurement (σ^2_{E}). Unfortunately, although all of the factors mentioned previously can be viewed as systematic error variance, only a few can be quantified by experimental or statistical procedures, such that a factor's specific effect on the gain scores can be estimated and removed from σ^2_{OG}.

Based on the many years of experience with Title I program evaluations and the invalidity issues examined in this section, there appear to be 11 factors that can increase pretest-posttest gain scores from September to June in any given school year:

1. history
2. maturation
3. statistical regression
4. small class size ($n < 30$)
5. overall school effects

6. test-wiseness
7. score conversion errors
8. "minor" variations in test administration
9. teaching to the test
10. coaching on test-taking skills
11. random error

A few studies of regression effect with classes composed primarily of low achievers (Linn, 1980a; Roberts, 1980; Tallmadge, 1982), small class size (Horst, 1981), score conversion errors (Elman, n.d.; Finley, 1981), and random error (Tallmadge, 1982) indicate that these factors alone could account cumulatively for as much as a half standard deviation in gain. The degree to which the other factors could spuriously inflate the average gain is difficult to assess. Furthermore, the impact of the 11 factors in one classroom can also be very different from the impact in other classrooms within the same school.

When these 11 factors are considered in conjunction with the 42 student, school, and test characteristics described previously, the net effect is to produce a sizable gain in the students' achievement which is independent of the teacher's performance or classroom instruction. The cumulative effect of the factors that positively bias estimated gain appears large enough to overstate the amount of teacher effect by a substantial margin. Currently, this "margin" cannot be determined exactly. As a consequence, it would be difficult to set a criterion for superior teacher performance that exceeds both normally expected gain and the gain due to the various sources of invalidity and error in each classroom.

Summary

The preceding analysis of achievement gain suggests eight conclusions in the context of teacher evaluation:

Measurement of Gain

1. The low reliability of gain scores and their negative correlation with pretest scores do not appear to be serious deficiencies of gain scores, as previously believed.
2. Low reliability does not necessarily mean lack of precision,

and the negative bias of the correlation is an artifact of measurement error on the estimation of the correlation.

3. Improved approaches to measuring gain supplement pretest-posttest data with between-person information.
4. The major limitation of the pretest-posttest gain score is the meager information it yields.
5. Multiwave data based on three, four, or more data points are preferable to two-wave data.

Validity of Gain-Score Inferences

6. The sources of invalidity of the pretest-posttest design include history, maturation, statistical regression, mortality, and interactions with selection.
7. There are 11 factors that can increase achievement gain.
8. The net effect of about 50 identified sources of invalidity is to produce a sizable gain in achievement that is independent of a teacher's performance.

CRITERION FOR SUPERIOR TEACHER PERFORMANCE

The career-ladder movement is designed to reward excellence in teaching. Ultimately, the incentive programs are intended to make the teaching profession more attractive in order to encourage the best and brightest to become and remain teachers (Southern Regional Education Board, 1986, p. 6). If excellence or outstanding teaching is the grounds for promotion and pay-bonus decisions, this standard for a teacher's performance must be expressed in concrete, operational language. If gains are to be used to identify the "superior teacher," then a criterion mean-gain score must be specified. What makes this task particularly difficult is the term *superior.* The implication is that the mean gain score of a class (or school) must be well above average or above the level of gain that could normally be expected from 10 months of teaching.

There are at least three major approaches one can pursue in an attempt to provide an operational definition for the criterion of superior teacher performance: (a) statistical significance, (b) educational significance, and (c) normative significance. The appropriateness and feasibility of these approaches are examined in this section.

Statistical Significance

One approach to assessing the degree of pretest-posttest achievement gain is to compute the *t*-test for paired observations. If the resulting *t* statistic reaches significance, it can be said that the gain is a "real" rather than a chance occurrence. Degree of gain is, therefore, defined as the magnitude of gain necessary to be found statistically significant.

Statistical significance is an unsatisfactory definition for two reasons. First, no graduated scale of gain is possible to differentiate normal from superior. Either a real gain is found or it is not. And second, because the power of a statistic is so dependent on sample size, teachers with relatively small classes would probably have insignificant gains and those with larger classes would have a better chance of obtaining significant gains. For example, for a class composed of 30 students, there would be greater than a 90% chance of attaining significance for a large gain; whereas for classes of between 10 and 20 students, there would be a 50% to 80% probability, respectively, of detecting similar gains (see Cohen, 1977, chap. 2).

All of these estimates of power could be decreased after considering the unreliability of the test(s). The appropriate pooled within-class reliability estimate for test-retest or parallel forms data has been developed by Subkoviak and Levin (1977, formula 3). Adjustments for unreliability are especially important in view of the fluctuation in power estimates for classroom size samples.

Educational Significance

The question remains as to just how much gain is indicative of superior teacher performance. One index that measures magnitude of gain is *effect size*. For pretest-posttest data, effect size is equal to the average gain score divided by the standard deviation of the test scores, assuming equal pretest and posttest variance (for details, see Cohen, 1977, chap. 2). Gain is simply expressed in standard-deviation units so that a magnitude of gain of, say, 0.5 or 1.0 standard deviation, can be specified as a standard for educational or practical significance. Criteria for what is deemed small, medium, and large gains can also be set.

Despite the availability of this meaningful index for defining "how much gain," determining the criterion for "superior" remains problematic. First, an analysis of class-by-class performances over several years would be required to ascertain the magnitude of gain that can normally be expected from 9 or 10 months of teaching. This analysis is complicated by the variability of class composition by grade level and subject area. Title I evaluation results, for example, suggest that marked differences in gain can occur between grades at the lower levels (Tallmadge, 1982). If it were found that a 0.5 standard deviation is a reasonable expectation for reading gain at a given grade level in a particular school, then at least a baseline has been established for setting a criterion for superior gain.

Second, one must wrestle with the multiple sources of invalidity and measurement error described in the preceding pages. It should be apparent by now that if a gain of 0.5 were found for a single class, it would be imperceptive to attribute that total gain to the teacher's performance. There are too many contaminating factors that could contribute to the estimate of gain. These factors must be addressed in order to isolate the amount of gain only due to in-class instruction.

Normative Significance

The statistical and educational significance criteria for superior teacher performance can be viewed as *absolute;* that is, a designated criterion can be met by one teacher irrespective of how other teachers perform. In fact, it is conceivable that no teacher may satisfy the criterion for "superior" at a particular point in time.

In contrast, the normative significance approach utilizes *relative* criteria, so that "superior" is defined in relation to a norm group of teachers. In one grade level at one school, for example, teachers may be ranked according to their estimated class gain scores. The teacher in the norm group with the largest gain may be identified as superior, relative to the other teachers in the norm group. The magnitude of gain necessary to be classified as superior may vary by grade level, subject area, and school. The implication is that *superior* has no absolute meaning as far as performance; it has relative meaning only.

Embedded within this relative meaning of superior are numerous sources of unfairness and inequity. Unless classes are comparable or matched on the factors discussed throughout this chapter, there are no defensible grounds for assuring a fair and equitable determination of superior performance. The between-class, between-grade, and between-student variability of the student, teacher, and test characteristics interacting with the sources of invalidity and error listed previously render any such determination as nearly impossible.

Summary

Three procedures for defining the criterion of superior teacher performance were examined. Statistical significance and educational significance provide absolute criteria based on probability and magnitude of gain, respectively. Normative significance establishes relative criteria, so that *superior* is defined in relation to a norm group of teachers. All of these approaches are unsatisfactory due to the problems inherent in defining *superior,* specific sources of bias (e.g., class size), and the multiple factors of invalidity and error that preclude the inference of superior teacher performance from achievement gain.

CONCLUSIONS

The four major sections of this chapter have described the difficulties one would encounter in developing a career-ladder or merit-pay program based on pretest-posttest student-achievement gain. These sections reviewed pertinent professional and legal standards, factors that influence a teacher's effectiveness beyond his or her control, the measurement and validity of gain, and, finally, approaches for determining the criterion of superior teacher performance. It is now possible to deduce several conclusions from the issues discussed:

1. There are no professional standards or court decisions to support the use of student-achievement data for any type of teacher evaluation.
   a. Standard 12.7 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1985) states that student test scores should not be used as the sole criterion for evaluating teachers or administrators.
   b. The *Personnel Evaluation Standards* do not recommend the use of student-performance data to evaluate teachers, administrators, or any other educational personnel.

   c. There are no standards that indicate student achievement should be one among several criteria for measuring teacher performance.
   d. All relevant technical standards, guidelines, and court decisions focus on the direct measurement of a teacher's performance.
2. An inference of superior, mediocre, or poor teacher performance from student achievement gains (or losses) can be contaminated by about 50 other factors.
   a. There are more than 40 student, school, and test characteristics that cannot be controlled by the teacher.
   b. There are 11 sources of invalidity of the pretest-posttest design that can increase achievement gain.
   c. The net effect of all of these factors is to produce a sizable gain in achievement that cannot be attributed to teacher performance or to classroom instruction.
3. Despite the traditional deficiencies of the low reliability of gain scores and their negative correlation with the pretest, the major limitation of gain scores is the meager information they provide based on only pretest and posttest measurements.
   a. Improved approaches to measuring gain supplement pretest-posttest data with between-person information.
   b. Multiwave data based on three, four, or more data points are preferable to two-wave data.
4. Between-class, between-grade, and between-student variability of the 50 sources of invalidity and error render the setting of a meaningful criterion for superior teacher performance nearly impossible.

Although there does not seem to be any single source of invalidity or error (systematic or random) that is large enough to invalidate the pretest-posttest gain-score model, the combination of multiple sources analyzed cumulatively does prove fatal to warrant its rejection as a primary strategy for measuring teacher performance in a career-ladder or merit-pay program. Even if student gains were to be considered as one among several evaluative criteria, the intractable problem of how they should be weighed in conjunction with other criteria must be tackled.

The professional standards, research evidence, and psychometric issues examined in this chapter strongly indicate that student performance on any test should not be used to measure teacher performance. Instead, that measurement should be guided by the *Personnel Evaluation Standards.* Teacher incentive programs should be designed according to those *Standards* and reflect the current state of measurement technology.

ACKNOWLEDGMENT

The author gratefully appreciates the helpful suggestions of John B. Willett and Kim Hoogeveen on an earlier version of this manuscript.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Bacharach, S. B., Lipsky, D. B., & Shedd, J. B. (1984). *Paying for better teaching: Merit pay and its alternatives.* Ithaca, NY: Organizational Analysis and Practice.

Benson, C. S., et al. (1965). *State and local fiscal relationships in public education in California.* Sacramento: Senate of the State of California.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 3–20). Madison: University of Wisconsin Press.

Berk, R. A. (Ed.). (1984a). *A guide to criterion-referenced test construction.* Baltimore: Johns Hopkins University Press.

Berk, R. A. (1984b). *Screening and diagnosis of children with learning disabilities.* Springfield, IL: Charles C Thomas.

Berk, R. A. (1984c, March). *The use of student achievement test scores as criteria for allocation of teacher merit pay.* Paper presented at the 1984 National Conference on Merit Pay for Teachers, Sarasota, FL.

Berk, R. A. (1986). Minimum competency testing: Status and potential. In B. S. Plake & J. C. Witt (Eds.), *The future of testing* (pp. 89–144). Hillsdale, NJ: Lawrence Erlbaum Associates.

Berk, R. A. (1988). Fifty reasons why student achievement gain does not mean teacher effectiveness. *Journal of Personnel Evaluation in Education, 1,* 345–363.

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work.* Boston: Kent–Wadsworth.

Bernardin, H. J., & Cascio, W. F. (1984). *An annotated bibliography of court cases relevant to employment decisions (1980–1983).* Boca Raton, FL: Florida Atlantic University.

Boardman, A. E., Davis, O. A., & Sanday, P. R. (1974). A simultaneous equations model of the educational process: The Coleman data revisited with an emphasis upon achievement. In *1973 Proceedings of the American Statistical Association, social statistics section.* Washington, DC: American Statistical Association.

Bower, R. (1982, March). *Matching standardized achievement test items to local curriculum objectives.* Symposium paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Bowles, S. S. (1970). Towards an educational production function. In W. L. Hansen (Ed.), *Education, income, and human capital.* New York: Columbia University Press.

Bridge, R. G., Judd, C. M., & Moock, P. R. (1979). *The determinants of educational outcomes.* Cambridge, MA: Ballinger.

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.

Brophy, J. E. (1973). Stability of teacher effectiveness. *American Educational Research Journal, 10,* 245–252.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101,* 147–159.

Calhoun, F. S., & Protheroe, N. J. (1983). *Merit pay plans for teachers: Status and description* (ERS Report No. 219–21684). Arlington, VA: Educational Research Service.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cascio, W. F., & Bernardin, H. J. (1981). Implications of performance appraisal litigation for personnel decisions. *Personnel Psychology, 34,* 211–216.

Centra, J. A., & Potter, D. A. (1980). School and teacher effects: An interrelational model. *Review of Educational Research, 50,* 273–291.

Cohen, D. K., & Murnane, R. J. (1985, Summer). The merits of merit pay. *The Public Interest, 80,* 3–30.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

Cohn, E., & Millman, S. D. (1975). *Input-output analysis in public education.* Cambridge, MA: Ballinger.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Corder-Bolz, C. R. (1978). The evaluation of change: New evidence. *Educational and Psychological Measurement, 38,* 959–976.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"—or should we? *Psychological Bulletin, 74,* 68–80.

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

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.

Elman, A. (n.d.). *Quality control in Title I: Manual versus computer conversions of test scores.* Palo Alto, CA: American Institutes for Research.

Evendon, E. S. (1918). *Teachers' salaries and salary schedules in the United States, 1918–19.* Washington, DC: National Education Association.

Finley, C. J. (1981, September). *What can state education agencies do to improve upon the quality of data collected from local education agencies?* Palo Alto, CA: American Institutes for Research.

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 695–763). Washington, DC: American Council on Education.

Glasman, N. S., & Biniaminov, I. (1981). Input-output analyses of schools. *Review of Educational Research, 51,* 509–539.

Gramenz, G. W., Johnson, R. C., & Jones, B. G. (1982, March). *An exploratory study of the concept of curriculum-referenced norms using the Stanford Achievement Test, sixth edition.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Green, D. R. (1983, April). *Content validity of standardized achievement tests and test curriculum overlap.* Symposium paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Haertel, E. (1986). The valid use of student performance measures for teacher evaluation. *Educational Evaluation and Policy Analysis, 8,* 45–60.

Hambleton, R. K., & Eignor, D. R. (1978). Guidelines for evaluating criterion-referenced tests and test manuals. *Journal of Educational Measurement, 15,* 321–327.

Hanushek, E. A. (1972). *Education and race: An analysis of the educational production process.* Lexington, MA: Lexington.

Hardy, R. (1984). Measuring instructional validity: A report of an instructional validity study for the Alabama High School Graduation Examination. *Journal of Educational Measurement, 21,* 291–301.

Hopkins, K. D., & Stanley, J. C. (1981). *Educational and psychological measurement and evaluation* (6th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Horst, D. P. (1976). *What's bad about grade equivalent scores, ESEA Title I evaluation and reporting system* (Tech. Rep. No. 1). Mountain View, CA: RMC Research Corporation.

Horst, D. P. (1981, March). *Title I evaluation and reporting system: Examination of the models at the project level.* Mountain View, CA: RMC Research Corporation.

Horst, D. P, Tallmadge, G. K., & Wood, C. T. (1974, October). *Measuring achievement gains in educational projects* (RMC Report UR–243). Los Altos, CA: RMC Research Corporation.

Jenkins, J. R., & Pany, D. (1978). Curriculum biases in reading achievement tests. *Journal of Reading Behavior, 10,* 345–357.

Joint Committee on Standards for Educational Evaluation. (1981). *Standards for evaluations of educational programs, projects, and materials.* New York: McGraw–Hill.

Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards: How to assess systems for evaluating educators.* Newbury Park, CA: Sage.

Kiesling, H. J. (1969). *The relationship of school input to public school performance in New York State.* Washington, DC: Office of Education, U.S. Department of Health, Education, and Welfare.

Kiesling, H. J. (1970). *The study of cost and quality of New York school districts: Final report.* Washington, DC: Office of Education, U.S. Department of Health, Education, and Welfare.

Kleiman, L. S., & Durham, R. L. (1981). Performance appraisal, promotion, and the courts: A critical review. *Personnel Psychology, 34,* 103–121.

Leinhardt, G. (1983). Overlap: Testing whether it is taught. In G. F. Madaus (Ed.), *The courts, validity, and minimum competency testing* (pp. 153–170). Hingham, MA: Kluwer–Nijhoff.

Leinhardt, G., & Seewald, A. M. (1981). Overlap: What's tested, what's taught? *Journal of Educational Measurement, 18,* 85–96.

Leinhardt, G., Zigmond, N., & Cooley, W. W. (1981). Reading instruction and its effects. *American Educational Research Journal, 18,* 343–361.

Levin, H. M. (1970). A new model of school effectiveness. In A. Mood (Ed.), *Do* *teachers make a difference?* Washington, DC: Office of Education, U.S. Department of Health, Education, and Welfare.

Linn, R. L. (1979). Validity of inferences based on the proposed Title I evaluation models. *Educational Evaluation and Policy Analysis, 1*, 23–32.

Linn, R. L. (1980a). Discussion: Regression toward the mean and the regression-effect bias. In G. Echternacht (Ed.), *New directions for testing and measurement (No.8)—Measurement aspects of Title I evaluations* (pp. 83–89). San Francisco: Jossey–Bass.

Linn, R. L. (1980b). Evaluation of Title I via the RMC models. In E. L. Baker & E. S. Quellmalz (Eds.), *Educational testing and evaluation: Design, analysis, and policy* (pp. 121–142). Beverly Hills: Sage Publications.

Linn, R. L. (1981). Measuring pretest-posttest performance changes. In R. A. Berk (Ed.), *Educational evaluation methodology: The state of the art* (pp. 84–109). Baltimore: Johns Hopkins University Press.

Linn, R. L., Dunbar, S. B., Harnisch, D. L., & Hastings, C. N. (1982). The validity of the Title I evaluation and reporting system. In E. R. House, S. Mathison, J. A. Pearson, & H. Preskill (Eds.), *Evaluation studies review annual* (Vol. 7, pp. 427–442). Beverly Hills: Sage Publications.

Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre and posttesting periods. *Review of Educational Research, 47*, 121–150.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16*, 421–437.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 21–38). Madison: University of Wisconsin Press.

Madaus, G. F. (1983). Minimum competency testing for certification: The evolution and evaluation of test validity. In G. F. Madaus (Ed.), *The courts, validity, and minimum competency testing* (pp. 21–61). Hingham, MA: Kluwer–Nijhoff.

Madaus, G. F., Airasian, P. W., Hambleton, R. K., Consalvo, R. W., & Orlandi, L. R. (1982). Development and application of criteria for screening commercial, standardized tests. *Educational Evaluation and Policy Analysis, 4*, 401–415.

Mayeske, G. W., & Beaton, A. E. (1975). *Special studies of our nation's students.* Washington, DC: Office of Education, U.S. Department of Health, Education, and Welfare.

Mayeske, G. W., et al. (1973). *A study of the achievement of our nation's students.* Washington, DC: Office of Education, U.S. Department of Health, Education and Welfare.

McClung, M. S. (1979). Competency testing programs: Legal and educational issues. *Fordham Law Review, 47*, 651–712.

Medley, D. M., Coker, H., & Soar, R. S. (1984). *Measurement-based evaluation of teacher performance: An empirical approach.* New York: Longman.

Michelson, S. (1970). The association of the teacher resourceness with children's characteristics. In A. Mood (Ed.), *Do teachers make a difference?* Washington, DC: Office of Education, U.S. Department of Health, Education, and Welfare.

Moore, B. C. (1984). *The effects of merit pay on selected secondary school teachers in terms of alienation and motivation.* Unpublished doctoral dissertation, Indiana State University.

Murnane, R. J. (1975). *The impact of school resources on the learning of inner city children.* Cambridge, MA: Ballinger.

Murnane, R. J., & Cohen, D. K. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard Educational Review, 56,* 1–17.

Nathan, B. R., & Cascio, W. F. (1986). Introduction. Technical and legal standards. In R. A. Berk (Ed.), *Performance assessment: Methods and applications* (pp. 1–50). Baltimore: Johns Hopkins University Press.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin, 88,* 622–637.

O'Connor, E. F. (1972). Extending classical test theory to the measurement of change. *Review of Educational Research, 42,* 73–98.

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin, 82,* 85–86.

Overall, J. E., & Woodward, J. A. (1976). Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. *Psychological Bulletin, 83,* 776–777.

Pencavel, J. H. (1977). Work effort, on-the-job screening, and alternative methods of remuneration. In R. Ehrenberg (Ed.), *Research in labor economics* (Vol. 1, pp. 225–258). Greenwich, CT: JAI Press.

Popham, W. J. (1983, April). *Issues in determining adequacy-of-preparation.* Symposium paper presented at the annual meeting of the American Educational Research Association, Montreal.

Porter, A. C., Schmidt, W. H., Floden, R. E., & Freeman, D. J. (1978). Practical significance in program evaluation. *American Educational Research Journal, 15,* 529–539.

Porwoll, P. J. (1979). *Merit pay for teachers.* Arlington, VA: Educational Research Service.

Poynor, L. (1978, April). *Instructional dimensions study: Data management procedures as exemplified by curriculum analysis.* Paper presented at the annual meeting of the American Educational Research Association, Toronto.

Richards, J. M., Jr. (1976). A simulation study comparing procedures for assessing individual educational growth. *Journal of Educational Psychology, 68,* 603–612.

Roberts, A. O. H. (1980). Regression toward the mean and the interval between test administrations. In G. Echternacht (Ed.), *New directions for testing and measurement (No. 8)—Measurement aspects of Title I evaluations* (pp. 59–82). San Francisco: Jossey–Bass.

Robinson, G. E. (1983). *Paying teachers for performance and productivity: Learning from experience.* Arlington, VA: Educational Research Service.

Robinson, G. E. (1984). *Incentive pay for teachers: An analysis of approaches.* Arlington, VA: Educational Research Service.

Rogosa, D. R. (1980). Comparisons of some procedures for analyzing longitudinal panel data. *Journal of Economics and Business, 32,* 136–151.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92,* 726–748.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement, 20,* 335–343.

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika, 50,* 203–228.

Rosenshine, B. (1970). The stability of teacher effects upon student achievement. *Review of Educational Research, 40,* 647–662.

Schmidt, W. H. (1983a). Content biases in achievement tests. *Journal of Educational Measurement, 20,* 165–178.

Schmidt, W. H. (1983b, April). *Methods of examining mismatch.* Symposium paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Schmidt, W. H., Porter, A. C., Schwille, J. R., Floden, R. E., & Freeman, D. J. (1983). Validity a variable: Can the same certification test be valid for all students? In G. F. Madaus (Ed.), *The courts, validity, and minimum competency testing* (pp. 133–151). Hingham, MA: Kluwer–Nijhoff.

Seiler, E. (1984). Piece rate vs. time rate: The effect of incentives on earnings. *Review of Economics and Statistics, 66,* 363–375.

Soar, R. S., & Soar, R. M. (1975). Classroom behavior, pupil characteristics and pupil growth for the school year and the summer. *JSAS Catalog of Selected Documents in Psychology, 5*(200), (ms no. 873).

Soar, R. S., & Soar, R. M. (1983). Context effects in the teaching-learning process. In D. C. Smith (Ed.), *Essential knowledge for beginning educators.* Washington, DC: American Association of Colleges for Teacher Education.

Southern Regional Education Board. (1986, December). 1986—Incentive programs for teachers and administrators: How are they doing? *Career Ladder Clearinghouse.*

Strike, K., & Bull, B. (1981). Fairness and the legal context of teacher evaluation. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 303–343). Beverly Hills: Sage Publications.

Subkoviak, M. J., & Levin, J. R. (1977). Fallibility of measurement and the power of a statistical test. *Journal of Educational Measurement, 14,* 47–52.

Summers, A. A., & Wolfe, B. L. (1977). Do schools make a difference? *American Economic Review, 67,* 639–652.

Tallmadge, G. K. (1982). An empirical assessment of norm-referenced evaluation methodology. *Journal of Educational Measurement, 19,* 97–112.

Tallmadge, G. K., & Wood, C. T. (1976). *User's guide: ESEA Title I evaluation and reporting system.* Mountain View, CA: RMC Research Corporation.

Tucker, L. R., Damarin, F., & Messick, S. (1966). A base-free measure of change. *Psychometrika, 31,* 457–473.

U.S. Equal Employment Opportunity Commission, U.S. Civil Service Commission, U.S. Department of Labor, & U.S. Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*(166), 38290–38309.

Webster, H., & Bereiter, C. (1963). The reliability of changes measured by mental test scores. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 39–59). Madison: University of Wisconsin Press.

Wiley, D. E. (1976). Another hour, another day: Quantity of schooling, a potent path for policy. In W. H. Sewell, D. L. Featherman, & R. M. Hauser (Eds.), *Schooling and achievement in American society.* New York: Academic Press.

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.

Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15, pp. 345–422). Washington, DC: American Educational Research Association.

Williams, T. B. (1980, April). *The distributions of NCE, percentile, and grade equiv-*

alent scores among twelve nationally standardized tests. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Winkler, D. R. (1975). Educational achievement and school peer group composition. *Journal of Human Resources, 10,* 189–205.

Wolf, R. M. (1977). *Achievement in America.* New York: Teachers College Press.

Yalow, E. S., & Popham, W. J. (1983). Content validity at the crossroads. *Educational Researcher, 12*(8), 10–14, 21.

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement, 19,* 149–154.