Merit Pay

MEMORANDUM

To: Albert Shanker
Robert Porter

From: Marilyn Rauth

Date: May 31, 1984

Ron Berk of Johns Hopkins University presented the attached paper at a conference on merit pay sponsored by the American Center for Management Development in March. In it he shows why student achievement test scores are an invalid means of evaluating individual teachers for purposes of merit pay or otherwise.

I'm having him develop a short booklet for us which reduces this to layman's language and could be used with school board members, state legislators and the general public. He's excited about the prospect and if money negotiations go well, this piece could be ready by the convention or before.

My conception is that it would begin by explaining the value of test scores in monitoring and assessing attainment of school system goals and then the invalidity of their use in evaluation of individuals.

MR/s
opeiu2aflcio

Attachment

THE USE OF STUDENT ACHIEVEMENT TEST SCORES AS CRITERIA FOR ALLOCATION OF TEACHER MERIT PAY

Ronald A. Berk
The Johns Hopkins University

Invited paper presented at the 1984 National Conference on Merit Pay for Teachers, Sarasota, FL, March 22-23, 1984.

Introduction

At present there are nine school districts in seven states (Arizona, New Hampshire, North Carolina, Oklahoma, South Dakota, Texas, Utah) that use 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) test scores serve as the only evidence of educational productivity.

Achievement test score gains by students are often preferred to administrator or supervisor ratings of performance because the measurement is preceived as more objective. This objectivity, however, is illusory. While student 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 scores of students are used to infer the productivity of teachers, that inference can be tenuous, inasmuch as the measurement of teacher productivity is obtained indirectly from student achievement. That is, teachers are not being measured directly.

The assessment of superior teacher performance or productivity in order to award merit pay 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 numerous difficulties. The difficulties stem primarily from limitations in testing technology coupled with the infeasibility of executing rigorous experimental design procedures in the natural school setting.

The purpose of this paper is to identify the major sources of difficulty in using student achievement test scores as criteria for allocation of teacher

merit pay. The sources are examined in the context of the most frequently used pretest-posttest design. It is, perhaps, the simplest, most efficient, and cost-effective model for estimating educational productivity. Only two test administrations are required: one at the beginning of the school year (September or October) and one at the end of the year (May or June). One test form or parallel forms can be used. The difference in class performance between the pretest and posttest is computed and the resulting "gain score" is used to infer teacher productivity.

The paper is organized according to six topics: (1) achievement test and score selection, (2) structure and content of instruction, (3) student characteristics, (4) reliability of test scores and gain scores, (5) validity of test score and gain score inferences, and (6) criterion for superior teacher productivity. Conclusions are given for the effectiveness of the pretest-posttest gain score model for awarding merit pay.

Achievement Test and Score Selection

Test Selection

Norm-referenced vs. criterion-referenced tests. The first decision that must be made is the type of achievement test(s) to be used to measure educational productivity. 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 Joint Technical Standards for Educational and Psychological Testing (AERA, APA, NCME Joint Committee, in preparation). 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 educational productivity.

Since standardized, norm-referenced tests such as the Iowa Tests of Basic Skills, California Achievement Tests, Comprehensive Tests of Basic Skills, Metropolitan Achievement Tests, and Stanford Achievement Test 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, 1983). 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. The assessment of curricular and instructional validity is described further in the section on validity.

In contrast to standardized tests, criterion-referenced competency tests are tailored to measure the instructional objectives of a school-based program (Berk, 1980, in press a). Such tests, however, must be developed by the local or state educational agency. 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, in press b). Commercially-developed criterion-referenced tests have also been plagued by numerous technical problems (Hambleton & Eignor, 1978).

One test vs. parallel forms. When the intervening period of time between testing is lengthy, say, three months or more, there is no statistical advantage to using a parallel test form on the second test administration. A parallel or equivalent form of the pretest, however, may be desirable for other reasons, especially to maintain test security. If a parallel form is

to be used, equivalence and equivalence-stability coefficients should be inspected. If the coefficients do not meet minimal standards, the parallel form should not be administered.

Score equating for parallel forms. If parallel forms are chosen, score equating is necessary. A parallel form's reliability coefficient provides evidence only of the degree of equivalence; when this equivalence is less than perfect, individual scores will differ on the two forms. For example, one form of a test, Form B, may be easier than another form, Form A. If no adjustment in the scores were made to account for those differences in difficulty, a score, of say, 60, on each form would mean something different. It would be harder to attain that score on Form A. If Form A was administered as the pretest and Form B was the posttest, an observed gain score could be very misleading. It would be attributable to the difficulty levels of the tests rather than to true achievement gain. The scores must be equated across Forms A and B to adjust for these differences and to establish their comparability for estimating gain scores.

This horizontal equating of test forms that are designed to measure the same content at the same level for the same population can be accomplished by using any one of four models: linear, equipercentile, one-parameter logistic Rasch), and three-parameter logistic (see Angoff, 1971; Holland & Rubin, 1982; Marco, 1981). The equating process transforms the raw scores on the two forms into one scale, often called scaled scores. Although there are systematic equating errors associated with these scores, they are typically less serious than the errors that can result from estimating gain scores from parallel forms which have not been equated.

Test Score Metric

In order to perform the most basic arithmetic calculations, such as computing the difference between pretest and posttest scores and group average scores, equal-interval scales are essential. The most popular derived score scale for norm-referenced tests is the grade equivalent. Unfortunately, it is not an interval scale and has several serious deficiencies (see Angoff, 1971; Berk, 1981, 1984, chap. 3; Cole, 1982; Flanagan, 1951; Horst, Tallmadge & Wood, 1974; Linn, 1981; Williams, 1980). Eight deficiencies have been identified by Berk (1984, chap. 3):

> 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. do not comprise an equal-interval scale;
> 7. exaggerate the significance of small differences in performance;
> 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)

Since 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. While 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 metrics for gain score analysis are standard scores such as z-scores, T-scores, and normal curve equivalents (NCEs), and scaled scores. They possess the desirable property of equal intervals and provide a common language for test to test, class to class, or other comparisons. Zimmerman and Williams (1982) stressed that the transformation of raw scores to one of the standard score metrics should occur after the individual difference scores have been computed. That is, the raw gain score, $X_2 - X_1$, should be calculated for each student first; then the gain score should be converted to the standard score scale. The authors indicate that this procedure is necessary so that the reliability of the raw gain scores is the same as the standardized gain scores. If the transformation is performed prior to determining the gains, "these gain scores can be utterly unreliable" (p. 153). When parallel test forms are employed, the (horizontally) scaled scores should be used.

For criterion-referenced tests the foregoing scores of relative standing are not meaningful. The simple proportion of items that a student answers correctly on each testing is an appropriate metric to estimate gain. Proportion correct is, in fact, an absolute as opposed to relative measure of achievement. Also, Linn (1981) has 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.

Structure and Content of Instruction

Objectives Guiding the Instruction

It is often desirable instructionally to state realistic instructional and behavioral objectives for each child along with appropriate prescriptions. This practice is required for all handicapped children according to the rules and regulations for implementation of P.L. 94-142 (U.S.O.E., 1977). The objectives and prescriptions are documented in the form of an individualized education program (IEP). Such within class variation among objectives and actual instruction, however, is inconsistent with the need to choose tests that measure some standard set of expected outcomes. The gap between this "standard set" and the "individualized set" can be sizable. The mismatch between the objectives the test measures and the objectives that actually guide the instruction can weaken the curricular validity of the test and the inferences from the gain scores (for a further discussion of curricular validity, see the section on validity).

Furthermore, the levels of cognition being taught in each classroom will frequently dictate the magnitude of individual gain scores. Knowledge level objectives and low level comprehension objectives requiring simple recall of factual content may exhibit large performance differences between pre- and posttestings. Impressive gains for these objectives should be anticipated. On the other hand, it would also be reasonable to expect that objectives designating complex concepts or skills at the upper levels of the cognitive hierarchy, (e.g., application, analysis) may not demonstrate pronounced changes in individual or group performance as a result of the specific instructional program. Gains for these objectives may not be observable for several months

or even years.

Since the content of the objectives specified by a teacher are determined by the student's instructional level(s), it is possible for greater achievement gains to result in the following types of classes: (a) those at the lower grade levels, (b) those where basic skills or knowledge objectives are being taught, and (c) those composed of low or underachieving students (e.g., learning disabled). Comparatively lower gains may be found for classes at the upper grade levels, for classes where higher level skills are stressed, and for classes composed of high ability students. Of course, the types of objectives assessed by the test and the heterogeneous composition of individual classes can markedly affect these trends.

Accessibility of Instructional Materials

The instructional materials and resources needed for teaching should be accessible to all teachers. If some teachers have constraints on what materials they can use in their classrooms and other teachers do not, instructional effectiveness can be impeded and gain score comparisons among teachers would be unfair. The methods employed by teachers to attain instructional objectives may vary; however, the materials required to execute those methods may not. At least, if the materials do vary from classroom to classroom, that variability should be due to teacher choice, not to administrator edict.

Student Characteristics

As suggested in the preceding sections, the composition of a given class can have an impact on pretest-posttest achievement gains. The specific direction of this impact will be governed primarily by the ability distribution in the class and the demographic characteristics of the students.

Ability

An ability distribution can be described as homogeneous or heterogeneous based on the amount of test score variance and also as high, average, or low based on arbitrary cutoffs above and below the mean. In general, a homogeneous class of average to high ability students who perform poorly on the pretest in September can demonstrate the most dramatic gains over a six- to nine-month period. These children have the greatest potential and chance for improvement in achievement test scores. On a set of basic skills objectives, these children have a high probability of exhibiting performance gains as a result of statistical regression effect (see section on validity), irrespective of classroom instruction.

Any other type of ability distribution will benefit to a lesser extent from noninstructional factors. In particular, underachieving students (a.k.a., learning disabled) have a lower overall potential for achievement gains and gifted students often have little possibility for improvement on many in-grade standardized tests where they have performed at or near the test ceiling on the pretest. Therefore, classes containing proportions of learning disabled and/or gifted students can be expected to yield average gain scores lower than classes without such students, where all other characteristics are similar.

Demographics

Socioeconomic level. Certain demographic characteristics of students also interact with achievement to produce either inflated or deflated gain scores. For example, the socioeconomic levels of students and their geographic location in the school district can influence accessibility to library facilities, an academic environment in which to study, microcomputers, and the like. Students

who are disadvantaged in relation to these educational supports may not manifest performance gains comparable with other students in other classes.

Sex. The proportion of males to females within a class can contribute to differential gains. Rate of learning to read and to solve mathematical problems differs for males and for females, especially in the primary grades. Whether achievement motivation or other factors can explain such differences is not clear. However, it is clear that classes composed predominantly of males will rarely yield average gain scores in reading and mathematics the same as classes composed mostly of females. Sex differences of students within and across classes should be considered in interpreting the educational productivity of teachers.

Reliability of Test Scores and Gain Scores

Test Scores

Reliability refers to the degree of consistency between two or more measurements of the same thing. It may be the individual scores or the decisions based on those scores that are analyzed over repeated measurements using a single test or parallel test forms. Among the different types of reliability that account for different sources of error in the scores, those most informative in the assessment of pretest-posttest achievement gains are internal consistency, test-retest (stability), parallel-forms (equivalence), and equivalence and stability.

If the same test is administered at the beginning and at the end of the school year, an estimate of internal consistency reliability such as coefficient alpha (or Kuder Richardson formula 20) should be computed for the test at each administration (r_{11} and r_{22}). It measures the adequacy of item sampling from the same content domain or item homogeneity; that is, the degree to which the items measure the same construct. High alpha

coefficients are desirable for estimating the reliability of gain scores.

In addition, the correlation between the scores from the two testings (r_{12}) should be calculated to furnish evidence of the stability of the scores over nine months. Publishers of norm-referenced tests usually report test-retest estimates for shorter time intervals (e.g., three to six months). A high coefficient of stability can reduce statistical regression effect between the pretest and posttest (see section on validity of gain scores), but at the time decreases the reliability of the gain scores.

When parallel test forms are administered, both estimates of equivalence and equivalence-stability should be obtained. Coefficients computed from the scores of parallel forms administered nine months apart will often be lower than test-retest coefficients because two sources of error are assessed: nonequivalence of item samples and instability of scores over time. These reliability coefficients must satisfy minimal standards for the intended score use. However, a high degree of equivalence between forms will not preclude the need to equate the scores on the two forms to conduct the gain score analysis.

Gain Scores

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

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

From this formula it can be seen that if the alpha coefficients are identical and equal to the test-retest coefficient, the reliability of the gain score is zero. Also, a high test-retest correlation tends to produce a low gain score reliability. Linn (1981, p. 87) gives an example for a test with a common variance and a reliability of .80. The reliability of the gain score would be .60, .50, .33, and .00 when the correlation (r_{12}) was .50, .60, .70, and .80, respectively.

While low reliability of gain scores is a serious concern in individual decision making, it is not a "fatal flaw" in group decision making where an average gain score is computed. This first deficiency is not an intractable problem in measuring teacher productivity.

The second deficiency of gain scores is their negative correlation with pretest scores. If the pretest and posttest 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).

A variety of methods has 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. Interestingly, the findings of investigations comparing these procedures (e.g., Corder-Bolz, 1978; Overall & Woodward, 1975, 1976; Richards, 1976) and the statistical models based on multiwave data (as opposed to two-wave or two-occasion pretest-posttest data) recently recommended by Rogosa and his colleagues (Rogosa, Brandt, Zimowski, 1982; Rogosa & Willett, 1983) and others (Nesselroade, Stigler, & Baltes, 1980) tend to diminish the seriousness of the aforementioned deficiencies. As 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.

In the measurement of educational productivity based on a pretest-posttest design, it can be argued persuasively that the simple mean difference score or raw gain between pre- and posttestings is about as accurate as any other estimate (Richards, 1976). However, a single inference derived from only two measurement points (e.g., September and May) can be regarded as questionable. Many different factors can invalidate an inference over such a lengthy time interval (see section on validity).

An alternative strategy for measuring gain that offers particular advantages over these pretest-posttest two-wave data is worthy of consideration. It involves the use of multiwave data (Rogosa et al., 1982). Multiple measurements such as September-January-May furnish additional information that can improve the precision of gain score estimates and the validity of productivity inferences drawn from those estimates.

Validity of Test Score and Gain Score Inferences

Test Scores

Validity is the degree to which a test achieves the purposes for which it was designed. It is inferred from the ways in which the test scores are used and interpreted. While content, criterion-related, and construct validity are applicable to achievement test scores in general, there are specific types of validity evidence that must be obtained to justify score inferences about teacher productivity. As mentioned in the earlier section on test selection, such evidence relates to curricular and instructional validity.

Curricular 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). While conceptually similar to content validity (Madaus, 1983; Schmidt, Porter, Schwille, Floden, & Freeman, 1983) and even viewed 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 (see, for example, Bower, 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 mismatch. 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 a lower level test that provides a better curricular match. Administering this below-grade-level test is called out-of-level testing. Its advantages and disadvantages have been discussed elsewhere (Arter, 1982; Berk, 1984, chap. 3).

Instructional validity. A concern related to curricular validity is whether standardized 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 (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) has identified four data-sources for describing whether students have received instruction that would enable them to perform satisfactorily on a test: (1) observations of classroom transactions, (2) analyses of instructional materials, (3) instructor self-reports, and (4) 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, so as to provide <u>adequate</u> evidence (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. Therefore, despite the importance of instructional validity, further research is required before it can be measured reliably, validly, and practically. (<u>Note</u>: The difficulties associated with gathering evidence of instructional validity are similar to those encountered in making fair and equitable decisions about merit pay.)

As demonstrated in the trial of Debra P. v. Turlington (1981), the foregoing types of validity evidence are applicable to criterion-referenced competency tests as well as standardized, norm-referenced tests (see also Hardy, 1983; Madaus, 1983). The Fifth Circuit Court ruled that "the state must demonstrate that the material on the test was actually taught in the state's [Florida] classrooms in order to establish the requisite 'content validity'" (Citron, 1982, p. 11).

Gain Scores

The validity of gain score uses 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 á 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 educational productivity will be described here.

The gain score computed from the pretest and posttest administrations is to be attributed to the teacher's effectiveness. The score is one indicant of his or her productivity. 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 student performance, only one of which may be teacher effort, then awarding that teacher merit pay 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 nine months between the testings which, in turn, affect student achievement. Home and community resources (e.g., books, computers) which may vary as 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 nine-month interval, 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 scorers 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. As noted in a previous section, highly spurious gains can occur for a class composed mainly of average to high ability students with poor pretest scores. These gains overestimate teacher productivity.

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 gain scores are compared across different classes in one school to determine which teacher deserves merit pay, there are additional factors such as selection-history, selection-maturation, and selection-instrumentation interactions that could account for differential gains in those classes.

Criterion for Superior Teacher Productivity

There are at least three major approaches one can pursue in an attempt to provide an operational definition for the criterion of superior teacher productivity: (1) statistical significance, (2) educational significance, and (3) normative significance. What makes this task particularly difficult is the term "superior." The implication is that the average gain score of a class must be well above average or above the level of gain that could normally be expected from nine months of instruction. The aforementioned approaches are examined from this perspective.

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 "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, since 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, .5 or 1 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 nine months of instruction. 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 .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 .5 were found for a single class, it would be imperceptive to attribute that total gain to the teacher's effectiveness. 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.

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

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

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, while 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}.

At present it is possible to determine the direction of the effect, increase (positive) or decrease (negative), for most of the factors. Based on the many years of experience with Title I program evaluations and the issues examined in this paper, there appear to be 13 factors that can <u>increase</u> pretest-posttest gain

scores from September to May in any given school year:

1. history
2. maturation
3. statistical regression
4. overall school effects
5. low level cognitive objectives
6. small class size ($n < 30$)
7. average to high ability levels
8. test-wiseness
9. score conversion errors
10. "minor" variations in test administration
11. teaching to the test
12. coaching on test-taking skills
13. 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, no date; 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 13 factors in one classroom can also be very different from the impact in other classrooms within the same school.

Those factors that can <u>decrease</u> measured gain include the degree of curricular and instructional invalidity of the test, high level cognitive objectives, and a high proportion of underachieving or gifted students. Factors for which the bias may be either positive or negative are mortality (or attrition) and test score

equating errors.

The net effect of these 19 different factors is to produce a sizable gain in achievement independent of teacher effort or instruction. The cumulative effect of the 13 factors that positively bias estimated gain appears large enough to overstate the amount of teacher effect by a substantial margin. Currently, this "margin" can not be determined exactly. As a consequence, it would be difficult to set a criterion for superior teacher productivity that exceeds both normally expected gain and the gain due to the various sources of invalidity and error in each classroom.

Normative Significance

The statistical and educational significance criteria for superior teacher productivity 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 that norm group with the largest gain may be identified operationally 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 productivity; 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 paper, there are no defensible grounds for assuring a fair and equitable determination of superior productivity. The between-class, between-grade, and between-subject variability of student characteristics interacting with the 19 sources of invalidity and error listed previously render any such determination as fallacious.

Conclusions

The various sections of this paper have described the difficulties one would encounter in developing a teacher merit pay system based on pretest-posttest class gain scores. The different stages of development were scrutinized, from the initial stage of achievement test selection through the specification of a criterion score for superior teacher productivity. It is now possible to deduce several conclusions about the process from the issues that emerged:

1. The pretest-posttest gain score model is afflicted with numerous sources of invalidity and measurement error.
2. Between-class, between-grade, and between-subject variability of objectives, instruction, resources, and student characteristics preclude (a) the trouble-free selection of an appropriate achievement test, (b) the precise estimation of gain, (c) the setting of a meaningful criterion for superior teacher productivity, and (d) the inference that estimated gain is attributable solely to teacher effort.
3. Although there does not seem to be any single source of invalidity or error (systematic or random) that is large enough to invalidate the model, the combination of multiple sources analyzed cumulatively does prove fatal to warrant rejection of the model.

4. Gain score evidence can be so misleading that it should not even be used to corroborate other evidence of teacher effectiveness or performance (e.g., administrator ratings).

Many of the obstacles in the path of the gain score model stem from its indirect measurement of teacher performance. Even if gain scores could be measured precisely, there still remains an inferential leap from improved student performance to superior teacher productivity. As the sources of invalidity strongly indicate, this leap is of nontrivial proportions.

Obviously it is premature to use achievement gain scores to infer superior teacher productivity as criteria for awarding merit pay. The measurement, statistical, and design issues examined in this paper render such a practice as indefensible. It would be exceedingly difficult, if not impossible, to logically, theoretically, or empirically justify the practice as fair and equitable for all teachers. Certainly if the gain score model is indefensible on these grounds, it will probably be indefensible on legal grounds as well.

References

AERA/APA/NCME Joint Committee. (in preparation). Joint technical 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.

Arter, J. A. (1982, March). Out-of-level versus in-level testing: When should we recommend each? Paper presented at the annual meeting of the American Educational Research Association, New York.

Berk, R. A. (Ed.). (1980). Criterion-referenced measurement: The state of the art. Baltimore, MD: Johns Hopkins University Press.

Berk, R. A. (1981). What's wrong with using grade-equivalent scores to identify LD children? Academic Therapy, 17, 133-140.

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

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

Berk, R. A. (in press b). Minimum competency testing: Past, present, and future. In B. S. Plake & J. C. Witt (Eds.), Future directions of testing and assessment. Hillsdale, NJ: Erlbaum.

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.

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.

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.

Citron, C. H. (1982). Competency testing: Emerging principles. Educational Measurement: Issues and Practice, 1, 10-11.

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

Cole, N. S. (1982, March). Grade equivalent scores: To GE or not to GE. Division D vice-presidential address presented at the annual meeting of the American Educational Research Association, New York.

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.

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

Debra P. v. Turlington, 644 F.2d. 397, 404 (5th Cir. 1981).

Elman, A. (no date). Quality control in Title I: Manual versus computer conversions of test scores. Palo Alto, CA: American Institutes for Research.

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.

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.

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

Hardy, R. (1983, April). Measuring instructional validity: A report of an instructional validity study for the Alabama High School Graduation Examination. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Holland, P. W., & Rubin, D. B. (Eds.) (1982). Test equating. New York: Academic Press.

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 (Technical Report 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.

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.

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, CA: 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, CA: Sage.

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, MD: 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. Pearsol, & H. Preskill (Eds.), Evaluation studies review annual (Vol. 7) (pp. 427-442). Beverly Hills, CA: Sage.

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. P., 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.

Marco, G. L. (1981). Equating tests in an era of test disclosure. In B. F. Green (Ed.), New directions for testing and measurement (No. 11) -- Issues in testing: Coaching, disclosure, and ethnic bias (pp. 105-122). San Francisco, CA: Jossey-Bass.

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

Mehrens, W. A. (1983, April). Inferences to which domain and why and implications of any mismatch. Symposium paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

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.

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.

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, CA: Jossey-Bass.

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.

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 as 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.

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

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. Office of Education (1977). Education of handicapped children: Implementation of Part B of the Education of the Handicapped Act. Federal Register, 42, 42474-42518.

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.

Williams, T. B. (1980, April). The distributions of NCE, percentile, and grade equivalent scores among twelve nationally standardized tests. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Yalow, E. S., & Popham, W. J. (1983). Content validity at the crossroads. Educational Researcher, 12, 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.