DOCUMENT RESUME

ED 259 460 EA 017 923

AUTHOR Coker, Homer
TITLE A Study of the Correlation between Principals' Ratings of Teacher Effectiveness and Pupil Growth.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Apr 85
GRANT NIE-G-82-0029
NOTE 41p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Gains; Achievement Rating; *Administrator Attitudes; *Correlation; Elementary Secondary Education; Opinions; *Principals; Teacher Administrator Relationship; *Teacher Effectiveness; Teacher Evaluation; Validity
IDENTIFIERS Georgia; Teacher Performance Assessment Instrument

ABSTRACT

This study was undertaken to assess the accuracy of principals' judgments of the effectiveness of the teachers they supervise. Each of 46 principals was asked to fill out a brief form judging the overall effectiveness of each of the teachers in his or her school. The form asked how effective the teacher was in performing three roles: (1) promoting academic goals, (2) promoting affective goals, and (3) performing other professional functions. Each principal's judgments of teachers of a single grade were intercorrelated with expected achievement gains of pupils of high, average, and low ability in the teachers' classes. Analytical procedures similar to those used in "meta-analyses" were used to examine the resulting large set of correlations. Findings revealed that the relationship between principals' judgments of teacher effectiveness and pupils' gains on achievement tests is very low. The factor most closely related to the magnitude of the correlation between principals' judgments and pupils' gains was the grade taught by the teachers rated. Other factors tested that were found not to be significantly related to the size of the correlations were pupil ability, subject taught, teacher role judged, and interactions between and among these factors. Tables and notes are included.(TE)

Homer Coker
Georgia State University

A STUDY OF THE CORRELATION BETWEEN PRINCIPALS' RATINGS OF TEACHER EFFECTIVENESS AND PUPIL GROWTH*

ABSTRACT

The study reported here was undertaken for the primary purpose of assessing the accuracy of principals' judgments or opinions of the effectiveness of teachers they supervise. By far the principal basis for personnel decisions about teachers is s rating of each teacher involved made by the teacher's principal or his or her assistant. Since, because of the well-known "halo effect," the principal's overall opinion of the teacher rated is a major determinant of the rating that teacher receives, the question whether these opinions are valid is an important question to ask.

Relatively few attempts have been made in the past to validate principals' judgments or ratings against measures of teacher effectiveness based on achievement gains of pupils in their classes; and those few attempts have consistently failed. The clear implication is that neither the judgments nor the ratings are accurate that many decisions based on them are wrong decisions. It seems time someone designed and conducted a new study of the problem, one that would give the principals' judgments every possible chance to prove themselves valid if indeed they are.

Design of the Study. The study that will be reported here differs from those that have gone before it in that instead of correlating judgments of teachers of different grades made by different principals with measures of teacher effectiveness, only judgments made of teachers of the same grade by the same principal were used. It also differs in that the 46 principals studied were asked to record their overall judgments rather than recording judgments on several characteristics on a multifactor rating scale. Finally, the procedure used to estimate teacher effectiveness was different from, and possibly more valid than, those used in past studies.

Each principal in the study was asked to fill out a brief form indicating how effective each of the teachers in his or her school whose effectiveness he or she felt capable of judging. The form asked how effective the teacher was in performing three "roles": (I) promoting academic goals, (II) promoting affective goals, and (III) performing other professional functions.

Each principal's judgments of teachers of a single grade were then intercorrelated with expected achievement gains of pupils of high, average, and low ability in the teachers' classes. Since the number of classes per grade tended to be small in most schools, this meant that any one validity estimate was highly

* A final report on grant NIE-G-82-0029 to the National Institute of Education. April 1985.

unstable because it was based on a very small group of teachers. (The average number of teachers per correlation in the 87 grade groups was, in fact, only 3.7.) The number of correlations estimated, on the other hand, was quite large. Twenty-four correlations were calculated for each principal and grade group so that the most of the mean correlations estimated were quite stable. Analytical procedures similar to those used in "meta-analyses" were used to examine this large set of correlations.

Findings. The mean correlation between a principal's judgment of the role I effectiveness of teachers (of the same grade and subject) and measured teacher effectiveness with pupils of average ability was only .20, and differences in the mean correlations for different principals were not statistically significant. There was, therefore, no reason to disagree with the conclusions of previous studies: that the relationship between principals' judgments of teacher effectiveness and how much their pupils gain on achievement tests is very low.

The factor most closely related to the magnitude of the correlation between the principals' judgments and pupil gains was the grade taught by the teachers rated. Other factors tested which were not found to be significantly related to the size of the correlations were pupil ability, subject taught, teacher role judged, and interactions between and among these factors.

Principals' Ratings. Because a substantial number of the schools in the study were located in Georgia, a unique opportunity arose to study principals' ratings. As part of the process of teacher certification, all beginning teachers in these schools were observed and rated by their principals (and two other raters) on the **TPAI** (Teacher Performance Assessment Instruments), a particularly well constructed rating scale. If and when the state department of education makes these ratings it will be possible to study the relationship between such ratings and principals' overall impressions of the effectiveness of the teachers rated, as well as to assess the validity of the ratings directly.

INTRODUCTION AND OVERVIEW

It is difficult to overstate the importance to public education of economical, accurate and practicable procedures for evaluating teachers. Effective operation of the educational enterprise (or any other) requires that all personnel be used efficiently. This in turn requires accurate and timely personnel decisions which depend on administrators' ability to distinguish more effective teachers from less effective ones quickly, economically, and (above all) accurately. As we shall see, what evidence there is indicates that such distinctions are not possible with the methods of teacher evaluation in current use.

The vast majority of personnel decisions made in education (like those in such other fields as business, industry and the military), are based on subjective judgments of employee competence made by immediate supervisors and recorded in the form of ratings. The validity of such ratings and the accuracy of decisions based on them depends very much on how good a judge of competence the rater happens to be.

The use of ratings can be defended only if we are willing to assume that the principal or other person who supervises teachers is an expert judge of teacher effectiveness, that most or all of his or her judgments are valid. That this is true is taken for granted; how expert any particular principal is, or principals in general are, is a question no one ever seems to ask.

A few studies which did ask this question were done some years ago. All of them reached the same conclusion: that the validity of a rating made by the average principal is near zero. The implication is clear: that ***teacher personnel decisions based on pure chance would be just about as accurate as decisions based on principals' ratings are!***

Since the most recent of these studies was done more than a quarter of a century ago, using methodology then available, now seems to be a good time to reopen the question, to do a new study. This report will describe such a study, a project in which we collected new data and applied a modern statistical design, one free from certain methodological limitations of the earlier studies.

Statement of the Problem. The main question this study was designed to answer is: *How valid are principals' judgments of teacher effectiveness?* Three related questions also investigated are: *Are some principals' judgments more valid than others'? What are some of the factors which affect the validity of principals' judgments?* and *How much effect do principals' overall judgments have on their ratings of teachers on multi-factor rating scales?*

Justification. Being able to distinguish more effective teachers from less effective ones is the key to bringing about those improvements in the education of children that depend on the quality of the teaching in the schools. The current concern with the competence of teachers and the demand for higher standards, merit pay plans and the like is new only in being noisier than a continuing concern on the part of the public and the professions as well. Its solution depends almost entirely on being able to evaluate teachers accurately, an ability whose lack neither the public nor most educators seem to suspect. The complete failure of past attempts to establish the validity of the ratings universally employed to accomplish this task makes it imperative to discontinue their use unless or until evidence of their validity is obtained.

It is just possible that the failures of previous attempts to validate principals' judgments were due in whole or in part to defects in the designs of the studies, that the judgments were valid but their validity was not detected. In any new study, therefore, it seemed important to take particular care to correct these defects and to give the principals' judgments every possible chance to prove themselves valid (if indeed they are). The study therefore involves some methodological innovations.

Sample. The sample of principals and teachers used in the study was drawn from elementary schools in the southeastern United States, a substantial number of which were located in Georgia. The sample used contained 46 principals and 322 teachers.

Methodological Innovations. The traditional approach to the problem of validating principals' ratings has been to correlate' ratings of teachers of various grades and in various schools made by their principals on one hand with measures of teacher effectiveness based on test scores of the pupils they teach on the other. The same basic approach was used in this study, but it was modified in two important respects. Each estimated correlation was based on a sample of teachers of the same grade in the same school. Because of this, no principal was asked to compare teachers of different grades, and validities of ratings made by different principals were estimated separately. Finally, the estimates of the effectiveness of all teachers of the same grade were based on gains of pupils of the same level of ability instead of on the average gain of all pupils in a teacher's class.

As part of the process of being certified competent to teach in that state, all beginning teachers in Georgia schools are rated by their principals (and two other raters) on the *TPAI* (Teacher Performance Assessment Instruments), a multi-factored behaviorally anchored rating scale. The willingness of the Georgia department of education to release these ratings to us

makes it possible to examine the relationship of ratings made with one of the most carefully constructed behaviorally anchored rating scales in existence to the raters' overall judgments of the effectiveness of the teachers being rated, as well as to gains of pupils in their classes.

REVIEW AND CRITIQUE OF RELATED LITERATURE

There would be no sense in repeating a study that had already been repeated several times with consistent findings unless there were some reason to expect a different result this time. In the following pages we propose to demonstrate that there is such a reason by briefly reviewing and discussing past research in the validity of principals' evaluations of teachers. In particular, we will point out some methodological problems with these studies, especially in the procedures used to derive measures of teacher effectiveness from measurements of pupils' gains on achievement tests, problems which will be avoided in this study. We will discuss studies of the validity of the TPAI separately. for reasons that will become apparent later.

Studies of the Validity of Principals' Ratings

The focus of interest here is not so much on the validity of principals' ratings of teachers as such as on the validity of the overall opinions principals form of the effectiveness of teachers being rated. It is our contention that the principal's overall impression of a teacher's effectiveness (often called "halo') is the principal determinant of his ratings of that teacher.

The Halo Effect. The multi-factor teacher rating scale seems to have become popular with educators around the year 1915 [1]. Instead of recording his overall judgment of the effectiveness of the teacher being evaluated, The principal (or other person) using such a rating scale records separate judgments of the status or level of the teacher being rated on a number of different characteristics, each of which is supposed to be related to teacher effectiveness. These separate ratings are then summed (or combined in some other way) to yield an overall indicator of the effectiveness of the teacher being rated.

The teacher rating scale was embraced enthusiastically and promptly by educators [2], and is still used almost everywhere to evaluate teacher competence, teacher performance, and teacher effectiveness as well.

The influence of the rater's general impression of the competence or effectiveness of the person being rated was recognized very early [3], and came to be known as the "halo effect." [4] The high intercorrelations typically found among ratings of the same teacher on widely disparate characteristics give evidence of the strength of this effect. the validity of the total or composite scores teachers get on a multifactor teacher rating scale may, and probably does, depend more on the validity of principals' overall judgments of the teachers than on

the degree to which they possess any of the characteristics or factors listed on the instrument.

Since researchers in the past have usually asked principals to record their judgments on multi-factor rating scales rather than as global judgments, research on the validity of principals' ratings provides the best information available about the validity of principals' judgments of teacher effectiveness.

Nine Studies. A search of the literature has turned up no more than nine published studies in which principals' ratings of teachers have been correlated with measures of gains in test scores of pupils in their classes. [5]

None of these studies was originally designed to test the validity of principals' ratings. The validity of the ratings seems to have been taken for granted by the researchers, who looked upon principals' ratings and measures of pupil gains as alternative "criteria of teacher effectiveness" with which measures of various other teacher characteristics could be correlated to find out whether they were related to teacher effectiveness. Before doing so the authors of each of the nine studies chose to intercorrelate these alternative criteria with each other.

All nine studies reached the same conclusion: that the correlation between principals' ratings and measures of teacher effects on pupils is close to zero. In other words, the average validity of principals' ratings is close to zero. Figure 1 quotes the conclusions stated by the author of each study verbatim. Such unanimity is rare in educational research.

Barr's Conjecture. In discussing their findings, Barr and his colleagues suggest that the validity of a principal's ratings may depend on who the principal is; that, even though the average validity in the population of principals is low, there may be some principa s who are better judges of teachers than most, and whose ratings are valid. If this were so, it would be important to identify these principals, to find out how they differed from the others, and to train these other principals to imitate them. This is one of the questions the present study attempted to answer.

CONCLUSIONS REACHED IN NINE STUDIES THAT ATTEMPTED TO RELATE PRINCIPALS' RATINGS OF TEACHERS TO MEASURED GAINS OF PUPILS IN THEIR CLASSES

1. Anderson, 1954: "... no appreciable relationships exist between rating criteria and pupil attainment criteria." (p. 67.)

2. Barr et al., 1935: "The observed coefficients of correlation between the measures of teaching ability and the three measures of gain in pupil achievement are uniformly low." (pp. 107-109.)

3. Brookover, W.B: "Employers' ratings of teaching ability are not related to pupil gains in information." (p. 205).

4. Gotham, R.E: "... the criterion of pupil change apparently measures something different from that measured by teacher ratings." (p.165).

5. Hellfritsch, A.G. "Teacher rating scales ... are only slightly related to the observed pupil growth." (p.199).

6. Jayne, C.D: ... supervisory ratings... seem to lack reliability and validity [as measures of pupil gain]. (p.133).

7. Jones, R.D. "Whatever pupil gain measures in relation to teaching ability it is not that emphasized in supervisory ratings." (p.98).

8. LaDuke, C.V. ...supervision ratings here provided are invalid [as predictors of pupil gain.] (p. 97).

9. Lins, L.J. "The three criteria... [pupil gain, pupil evaluations of the teacher, and a composite of five supervisory ratings] are ... related to a greater degree than can be attributed to chance." (p.59).

10. Medley and Mitzel, 1959: "The results of the present study ... suggest that supervisory ratings do not correlate with [pupil] growth..." (p.244).

FIGURE 1

Procedures for Estimating Teacher Effectiveness

It seems clear from the foregoing iscussion that the correlation between principals' judgments of teacher effectiveness and measures of teacher effectiveness based on measured achievement gains of pupils tend to be very low. It is natural to attribute this low relationship to difficulties principals have in distinguishing more effective teachers from less effective ones; that is, to say that the judgments are not valid. But it is certainly possible that the low correlations may be due, at least in part, to defects in the measures of teacher effectiveness, that they lack validity. Let us consider this possibility.

Validity of Direct Measures of Teacher Effectiveness. The validity of a measure of a direct measure of teacher effectiveness, that is, one based on pupil gains on achievement tests, depends on two things: it depends first of all on the validity of the test or tests used as measures of achievement of the objectives the teacher is or ought to be working toward; and, second, on the degree to which it succeeds in isolating that part of the gains pupils make that results from the efforts of the teacher from that which would have taken place anyhow.

Let us begin by assuming tha the tests administered to the pupils are valid measures of objectives the teacher is expected to achieve. This assumption has been questioned by some on the grounds that the content of the items on the test may not coincide exactly with the items of content the teacher actually teaches. Our reasons for rejecting this notion will be given later. The assumption seems reasonable enough when, as is the case in this study, the test used is one adopted by the local school system as an appropriate measure of system-wide goals.

Isolating the Teacher's Contribution. Meeting the second condition is more difficult. If it were possible to assign pupils to classes randomly, so that at the beginning of the school year the classes taught by different teachers would differ only by .hance, there would be no problem. Any differences in post-test scores of pupils in different teachers' classes beyond those attributable to chance could safely be attributed to differences in the effectiveness of the teachers of those classes. But when pupils are not randomly assigned to classes, the classes differ at the beginning of the year in unknown ways and to an unknown degree. It is therefore necessary to distinguish among the differences found at the end of the year those that merely reflect differences that existed at the beginning of the year from those that did not, and somehow measure the latter in isolation from the former.

In past studies of teacher effectiveness, the basic approach to

this problem has been to estimate the mean achievement gain of all of the pupils in each teacher's class. and then to compensate for differences between the pupils in different classes by statistical adjustments.

We will introduce and use a different approach entirely. But before doing so, let us briefly review and comment on the most common procedures used in the past. All of them begin by regressing posttest scores on pretest scores and predicting the mean posttest score in each teacher's class with the regression equation. The difference between the mean of the posttest scores tne pupils in a teacher's class actually earn and mean of their predicted posttest scores is used as a measure of that teacher's effectiveness.

Residual Gains. The main differences in the three techniques that have been used is in how the regression line is estimated. In the earliest method, called the *residual gains* method, the regression was estimated from the variance and covariance between classes; that is, by intercorrelating class mean pretest scores wit class mean posttest scores. Mitzel and Gross, in their classic paper on the topic [6] reject this procedure on the grounds that it adjusts out some of the differences between classes that it is supposed to estimate.

Adjusted Mean Gains. Mitzel and Gross recommended, instead, the use of *adjusted mean gains*, that is, that the regression be estimated from pooled within-class variance and covariance.

Multiple Regression. More recently, some investigators have used total variance and covariance in a multiple regression in which pretest scores are entered first, then the variables with which teacher effectiveness is to be correlated.

Unfulfilled Assumptions. Use of any of these techniques is based on two assumptions that are rarely if ever fulfilled in practice. One is that the pupils have been randomly assigned to the classes of the different teachers; the other is that the regression slopes with classes are equal. As we have already noted, random assignment of pupils rarely happens. It is, of course, impossible unless the sample of teacher studied consists of teachers of the same grade and subject in the same school, because pupils cannot be assigned to grades, subjects, or schools at random.

The assumption that regression slopes (and therefore pretest-posttest correlations) within classes are equal is testable; and when it is tested is usually found to be false. The correlation between pretest and post test scores within a teacher's class is, in fact, a characteristic of the teacher that is important in its own right, since it reflects the degree to which the effectiveness of the teacher varies with pupil

ability. A positive slope indicates a class in which high ability pupils gain more rapidly than low ability pupils do; a zero slope indicates a class in which all pupils gain at the same rate, etc.

Fitting a single regression line to the pupils in different classes not only results in a poor fit, then, but it also conceals important information about teacher effectiveness.

Regression Artifact. The most defensible of these three procedures is, of course, the analysis of covariance, which does not confound between-class and within-class covariation. This procedure has also been widely used in quasi-experimental or ex post facto studies, ones in which subjects are nor randomly assigned to treatments, to achieve the same purpose, that is, to compensate for pre-existing differences between groups.

It has been shown, however, that because of an artifact of regression, [7] when this procedure is used with groups that differ initially it has the opposite effect to the one intended. That is, it increases the bias it is supposed to reduce.

What is important to us is that, since all nine of the studies cited earlier used procedures of this type, it is possible that a bias in the estimates of teacher effectiveness may have concealed the validity of principals' ratings in all of these them. To avoid this possibility, in the present study we will use a procedure different from any of those described, one which avoids both of the untenable assumptions implied in the use of the procedures described above.

Validity of the TPAI

Among many attempts to control or eliminate the halo effect, one of the most promising has been the use of "behavior anchors" on the separate scales of a multifactor rating scale. A behavior anchor consists of one or more specific examples of behaviors typical of teachers at a specific level on the dimension the scale is intended to measure. Their inclusion is intended to increase the accuracy of ratings on a subscale by clarifying and simplifying the task of the rater.[8]

Of special interest in this investigation is the carefully constructed behaviorally anchored rating scale (or set of scales) called the *Teacher Performance Assessment Instruments (TPAI)*. The TPAI was developed, and for several years has been used, for certifying beginning teachers in the state of Georgia.[9]

Studies of Validity of the TPAI. A series of studies of the predictive validity of the TPAI has been reported at various

research meetings; we propose to review these studies here.[10] Because they mainly report correlations between pupil gains and scores on individual TPAI items or competencies instead of total scores, results of these studies do not shed as much direct light on the question addressed by the present investigation as we might wish. They do not tell us as much about the validity of principals' judgments of the effectiveness of teachers as the nine studies already discussed. But they do bear directly on the questions about the accuracy of decisions about educational personnel with which this study is concerned.

Measure of Teacher Effectiveness. Three kinds of tests have been used in these studies to measure teacher effectiveness: standardized tests, teacher-made tests, and criterion referenced tests. By and large the correlations reported are correlations between measures of teacher effectiveness and scores on single TPAI items or competencies rather than total scores. The results obtained seem to depend on the kind of test used. When standardized tests were used, the correlations obtained are described by the authors as "mixed." When criterion-referenced tests are used at least some of the correlations reported tend to be significant. And when teacher-made tests are used, many more correlations are significant.

Test Content and the Nature of Effective Teaching. These authors raise a familiar objection to the use of standardized test scores of pupils to estimate teacher effectiveness, the objection that because a standardized test may not measure the exact content taught by the teacher, it is not a valid basis for assessing teacher effectiveness. This fallacy reflects a basic misunderstanding of the proper function of standardized tests, of the nature of effective teaching and, indeed, of the purpose of education.

It is the function of a standardized test to measure the important, permanent changes in pupils that teacher-made unit tests cannot measure. Growth in the ability to read critically, to apply the scientific method, to learn on one's own, and the like, is gradual, difficult to measure, and in most cases can be detected only over relatively long periods of time. These are the kinds of things teachers are hired to teach. These are the the kinds of outcomes that distinguish truly effective teachers from the rest. These are the kinds of outcomes on which measures of teacher effectiveness should be based.

Standardized tests are not, or should not, be designed to measure pupils' mastery of the specific content of the day-to-day lessons or units taught in the schools. This is what the unit test, which is usually built by the teacher, is supposed to measure. Most of it will be forgotten by the pupils promptly once they have passed the unit test.

The specific content that a teacher teaches in a lesson or unit is a means to the ends the teacher is supposed to achieve, but not the end itself. Content objectives are no more than "enabling" objectives; the actual content taught in a unit is not important and will soon be forgotten by most pupils, and rightly so. But in the process of learning (and forgetting) this content the teachers' pupils ought to learn something else which they will not forget, something only one of the better standardized tests can measure.

The unit tests that a teacher constructs to measure how much of the content of the unit pupils have learned are useful for such purposes as guiding and motivating pupils to learn the content, and providing a practical basis for giving them grades. How well a pupil learns the content is a pretty good indicator of how much progress the pupils is making toward the important goals of education.

So far as we know, none of the criterion-referenced tests so much in vogue these days are designed measure anything more than the specific content teachers are supposed to teach. It is important that the content of a criterion-referenced test matches that taught by a teacher. But pupil gains on such tests do not validly indicate how effective a teacher is in performing the basic function of a teacher, which is to *educate* children, to change them permanently and in important ways.

Only a standardized test, and a good one at that, is capable of measuring how successful a teacher is in educating pupils, and it can only do so by measuring changes over a substantial period of time, preferably a full school year. And even the best standardized test cannot do this when the teachers "teach to the test," that is, when they teach the specific content of the test. When that happens, the validity of the test as a measure of the important outcomes of education is destroyed; and it becomes, in effect, nothing more than another criterion-referenced test.

This is one concern we have with the TPAI validity studies: that they fail whenever standardized tests are used to assess teacher effectiveness, and succeed when tests that measure only the pupils' immediate mastery of content are used. But we have a more serious problem than that.

The Comparability Problem. Unless the same standardized test is administered to all classes in a study, The comparability of scores from different classes are not comparable unless something is done to make scores on different tests equivalent. The authors' solution to this problem was to use a statistic called the *Index of Achievement Gain*, which seems to be home-grown. A pupil's Index of Achievement Gain is calculated by dividing the increase in the number of items the pupil answers correctly from

the pretest to the posttest (the actual gain) by the number of items the pupil failed to answer correctly on the pretest (the possible gain). The mean of these indices for all pupils in a teacher's class was the teacher effectiveness measure used with teacher-made tests and criterion-referenced tests.[11]

There is no reason to suppose that this statistic yields comparable scores from non-comparable tests. Suppose, for example, that Miss Jones' slow- learning fifth grade pupils gain 10% on her 25-item unit test on improper fractions; and that Miss Smith's above- average pupils gain 15% on her unit test on the Civil war. On what basis can we conclude, as these investigators do, that Miss Smith is a more effective teacher than Miss Jones?

It is puzzling and disturbing to note that it is only when these investigators use this highly questionable statistic that they get significant correlations with TPAI scores. Whatever it is that indices of achievement gain based on non-equivalent tests measure, it is not the relative effectiveness of the teachers who built the tests.

It is more likely that these indices tell us something about the teachers' skill in constructing tests; but why should that correlate with scores on TPAI items? Can it be that whatever makes some teachers impress observers most favorably also makes them write test items on which their pupils make large percentage gains? Far-fetched as this explanation may be, it is more credible than the idea that these indices yield comparable measures of teacher effectiveness.

Perhaps the best conclusion we can reach about the validity of the TPAI as a measure of teacher effectiveness is that the question is still open. The fact that TPAI scores are used as at least a partial basis for deciding whether candidates will or will not be granted teaching certificates makes it worth while to try once more to validate it.

Summary and Conclusions

The facts that emerge from this brief look at the literature clearly call into question the wisdom of the almost complete dependence of personnel decisions in education on principals' ratings. The fact is that all attempts to establish the validity of such ratings against criteria of teacher effectiveness based on measured achievement gains of pupils have been unsuccessful. The validity of the methods used in these studies to estimate teacher effectiveness are, however, open to question. Until the possibility that methodological shortcomings may account for these findings can be ruled out, however, there is a need for studies which are free from these methodological flaws.

NOTES

1. Boyce, A.C. Methods of measuring t achers' efficiency. Fourteenth Yearbook of the National Society for the Study of Education. Part II. Bloomington, Ill: Public School Publishing Co. 1915.

2. Barr, A.S. What qualities are prerequisite to success in Teaching? The Nation's Schools, 1930, 6, 60-64.

3. Wells, F.L. A Statistical study of literary merit. *Archives of Psychology No. 7*, 1907.

4. Rugg, H.O. Is the rating of human character practicable? *Journal of Educational Psychology*, 1921. 12, 425-438,485-501; 1922, 13, 30-42, 81-93.

5. The nine studies are: Anderson, H.M. A Study of certain criteria of teaching effectiveness. *Journal of Experimental Education*, 1954. 23, 41-71; Barr, A.S., Torgerson, T.L., Johnson, C.E., Lyon, V.E. and Walvoord, A.C. The Validity of certain instruments employed in the measurement of teaching ability. Chapter IV in *The Measurement of Teaching Efficiency*, H.M. Walker, editor. New York, MacMillan, 1935; Brookover, W.B. The relation of social factors to teaching ability. *Journal of Experimental Education*, 1945, 13, 191-205; Gotham, R.E. Personality and teaching efficiency. *Journal of Experimental Education*, 1945, 14, 157-165; Hellfritsch, A.G. A Factor analysis of teacher abilities. *Journal of Experimental Education*, 1945, 14, 166-199; Jayne, C.D. A Study of the relationship between teaching procedures and educational outcomes. *Journal of Experimental Education*, 1945, 14, pp. 101-134;Jones, R.D. The Prediction of teaching efficiency from objective measures. *Journal of Experimental Education*, 1946, 15, 85-99; LaDuke, C.V. The Measurement of teaching ability. *Journal of Experimental Education*, 1945, 14, 75-100; Lins, L.J. The Prediction of teaching efficiency. *Journal of Experimental Education*, 1946, 15, 2-60; Medley, D.M. and Mitzel, H.E. Some behavioral correlates of teacher effectiveness, *Journal of Educational Psychology*, 1959, 50, 239-246.

6. Mitzel, H.E. and Gross, C.F. The Development of pupil-growth criteria in studies of teacher effectiveness. **Educational Research Bulletin**, 1958, 37, 178-187, 205-215.

7. Campbell, D.T. and Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make

compensatory education look harmful. in Hellmuth, J., editor, *The Disadvantaged Child Volume 3*, New York, Breimer/Mazel, 1971.

8. Cf. Ryans, D.G. *Characteristics of Teachers*, Washington, D.C., American Council on Education, 1960.

9. Capie, W., Anderson, S.J., Johnson, C.E., and Ellett, C.D. *Teacher Performance Assessment Instruments: A Handbook for Interpretation.* Athens, Georgia, Teacher Assessment Project, College of Education, University of Georgia, 1979.

10. These studies are summarized in: Capie, W. and Ellett, C.D. Issues in the measurement of teacher competencies: Validity, reliability and practicality of Georgia's assessment program. Paper presented at the annual meeting of the American Educational Research Association, New York, 1982.

11. Ellett, C.D., Capie, W. and Johnson, C.E. Teacher performance and pupil achievement on teacher-made tests. Paper presented at the annual meeting of the Eastern Educational Research Association, Norfolk, Virginia, 1980.

PROCEDURES

In this section of this report we will describe the selection of the sample of principals studied, the collection of the data, the instrumentation, the measure of teacher effectiveness, and the analytical methodology of the study.

The Sample

The sample of principals, teachers, and pupils used was obtained by seeking the cooperation of school districts in the southeastern United States. If a school district agreed to take part in the study, the next step was to find out whether the regular testing program in the district yielded the data needed in the study. If it did, each elementary-school principal in the district was asked to record judgments of the effectiveness of as many of the teachers in his or her school as possible. Usable data were obtained from 46 principals on 322 teachers.

Data Collection

Each principal in the sample recorded his or her judgment of the effectiveness of each teacher he or she supervised on a simple form. A roster of each class was obtained that showed the fall and spring scores of each pupil in that class on whatever test battery was used in the regular testing program in the district. The state education department of the state of Georgia kindly consented to provide us with ratings of any of our 322 teachers who were first-year teachers in the state of Ge a that they had obtained (although they have not yet done so).

Instrumentation

Three instruments were used in the study: the form on which principals recorded their judgments of teachers, the achievement tests administered to the pyupils in the 322 classes, and the rating scale used in the Georgia certification program.

Principlas' Judgments. The instrument on which the principals were asked to record their judgments was a simple form used in a study reported in 1959. [1] (See Figure 2.) On it the principal indicates where the teacher would stand in comparison with a typical group of 20 teachers of the same grade on three "roles" a teacher is expected to perform, defined in Figure 2.

INSTRUCTIONS TO PRINCIPALS

Teachers in today's schools must perform competently in at least three roles in order to be successful. You are being asked to share with us your best judgment as to how well the teacher named above fulfills each of them in your school as a teacher of the subject named.

Please indicate your judgment by writing a number between one and twenty in the space before the description of each role printed below. The number should indicate where you think the teacher would rank in a representative group of teachers in that subject and grade. If the teacher performs better than all the rest, write 20; if all the others perform better than this teacher, write 1; and so on.

All ratings will be kept confidential; no one except the clerk who transcribes the data (and removes all names) will know the name of either the teacher or the principal involved. These sheets will be destroyed as soon as the data have been transcribed.

___ROLE I The teacher is responsible for providing learning experiences which result in pupils' acquisition of fundamental knowledge.

___ROLE II The teacher is responsible for providing children with learning experiences which lead to good citizenship, personal satisfaction, and self understanding.

___ROLE III The teacher is a professional colleague of other teachers, supervisors, and administrators.

FORM ON WHICH PRINCIPALS RECORDED THEIR JUDGMENTS OF TEACHER EFFECTIVENESS

FIGURE 2]

Achievement Tests. In each school, the reading and arithmetic subtests of the battery used in the regular testing program in the school were used to measure achievement gains of pupils.

In the conventional study of teacher effectiveness, in which teacher ratings and teacher effectiveness measures are intercorrelated across schools, it is necessary to use the same tests in all classes so that the teacher effectiveness measures are comparable. But since all correlations in this study were calculated in groups of teachers of the same grade in the same school, it was not necessary to use the same test in every school. Instead, the test used in each school was the one chosen by that school as most appropriate. When we asked a principal how effective a teacher was, we meant how effective in terms of a test already in use in that school with which both the principal and the teacher were already familiar, and one which presumably measured the goals of the school.

Rating Scale. The rating scale used in the second phase of the study was the TPAI *Teacher Performance Assessment Instruments*, which was developed and is used in Georgia as one of a number of instruments used as a basis for certifying teachers in the state. It was chosen mainly for the reason already given; that ratings made of beginning teachers were on file and available. It would have been an excellent choice in any case since it is one of the most carefully constructed and widely used behaviorally anchored multi-factor rating scales in existence.

Expected Gain Scores

The measure of the effectiveness of each teacher that was used in this study was the **Expected Gain Score** of a pupil with a specified level of ability as indicated by his or her pretest score on the test used to measure achievement gains. Since this measure has never to our knowledge been used before for this purpose, we propose to describe it here in some detail.

A pupil's Expected Gain Score or EGS is an estimate of the score he or she will earn at the end of the school year; it depends, among other things, on the pupil's ability and on which teacher's class he or she is in. Differences between scores the same pupil would be expected to get in different teachers' classes will be used as measures of differences in teacher effectiveness.

How is the score the pupil will get at the end of a school year in a teacher's class (his or her EGS) estimated? By entering the pupil's pretest score into a simple linear regression equation based on the correlation between the pretest and posttest scores of all of the pupils in that teacher's class. Such a regression

equation looks like this:

$$y = a + bx$$

The values of a and b, known as the *regression coefficients*, depend on which class the pupil is in. The value of x, the pretest score, can be anything you choose within the range of scores on the test. From these three numbers the value of y, the EGS, can be calculated. Since the values of the regression coefficients a and b will differ from one teacher's class to another, the EGS score obtained with any given pretest score will differ for different teachers. In other words, pupils with identical pretest scores will get different EGS's, will learn different amounts, in different teacher's classes.

The actual posttest score that any individual pupil with a given pretest score gets at the end of the school year may or may not equal the predicted posttest score or EGS; pupils with the same pretest score will differ in other ways that affect the amount they learn. But the average posttest score of a large number of pupils with that pretest score scores would equal the predicted value, the EGS. In other words, the EGS is an estimate of the mean posttest score in a population of pupils with the same pretest score.

While any arbitrarily chosen pretest score may be used, the average pretest score in some specific group is of greatest interest in most cases. Suppose, for example, that the mean score of all fifth-grade pupils in a school district on the pretest is substituted in a regression equation obtained in Miss Jones' fifth-grade class and in Miss Smith's fifth-grade class. Suppose that the EGS obtained in Miss Jones' class is 54 and that obtained in Miss Smith's class is 47. This indicates that the average pupil in that school system would gain 7 points more in Miss Jones' class than in Miss Smith's. Within the limits of the errors of measurement, we are justified in concluding that Miss Jones is more effective with the average pupil than Miss Smith.

In general, the teacher in whose class a pupil with a particular ability level (as measured on the pretest) would get the highest posttest score will be regarded as the teacher who is most effective with pupils at that ability level.Such EGS's are comparable for teachers in the same grade because the pretest scores are identical for all teachers. They are not usually comparable for teachers of different grades, however, because the average pretest score will differ for different grades.

Pupil Ability and Teacher Effectiveness. Some of the research suggests that which pattern of classroom behavior is most effective in promoting pupil gains in achievement depends on the

ability of the pupil [2]. If this is so, then one cannot assume that a teacher who is most effective with one type of pupil, such as the average pupil in a school district, is necessarily the most effective with all kinds of pupils.

This possibility has many important and disturbing implications. One is that it does not make much sense to ask a principal to judge the effectiveness of a teacher without specifying the kind of pupil to be affected. It might be that one principal bases his judgments on how effective a teacher is with low-ability pupils while the researcher was measuring how effective each teacher is with pupils of average ability.

For this reason, we estimated not one but two EGS's for each teacher, one for pupils whose pretest score is one standard deviation *below* the mean of the distribution of all pupils in the grade and school, and one for pupils whose pretest score is one standard deviation *above* the mean of the same distribution. The first pretest score was at the 16th percentile and the second at the 84th percentile of the distribution. so the first group of pupils will be referred to as "low-ability" pupils and the second as "high-ability" pupils. Because the regression is linear, the mean of these two EGS's is the EGS of pupils of average ability.

To sum up, then, we had three measures of the effectiveness of each teacher; one with low-ability pupils, one with high-ability pupils, and one with pupils of average ability. A correlation between a principal's judgments and any one of these will be interpreted as evidence that his judgments are valid.

As an measure of teacher effectiveness, an EGS score is subject to measurement error. In order to obtain an estimate of this error, we split each teacher's class into random halves and calculated not one but two regression equations per class, one from each half. Substituting the same pretest score in each equation gave us two independent estimates of the same EGS. The mean of the two was used as the estimate of teacher effectiveness with pupils of the level of ability in question, and the difference between the two half-class values was an indicator of its accuracy.

Thus there were four expected gain scores per class for each test, two for high-ability pupils (one in reading and one in arithmetic) and two for low-ability pupils, making eight in all. Each of these these eight expected gain scores was correlated with principals' judgments of the effectiveness in performing each of the three roles of the teachers in each grade in each school, yielding 24 correlations per grade group. If a principal recorded judgments on teachers in 6 grades in his or her school, then we calculated 24X6 coefficients for that principal.

Data Analysis

The first step in the analysis of the principals' judgments of teacher effectiveness in the three roles was to calculate a set of validity coefficients (correlations between principals' judgments and expected pupil gains) for each principal. It did not seem reasonable to us to compare judgments of teachers of different grades, so separate validity coefficients were calculated for the teachers of each grade who were judged by the same principal. Correlations were calculated between judgments on each of the three roles and EGS's of pupils of high and low ability, in reading and arithmetic, in random half-classes, making a total of 24 correlations for each grade judged by each principal, as well as mean correlations for grades, subjects, etc.

Because the number of classes per grade in a school tended to be small, as it is in most schools, most of these correlations were based on rather small groups of teachers. The average number of teachers in one grade group was, in fact, only about 3.7. The number of correlations estimated for each principal, on the other hand, tended to be quite large, so that the mean correlation between a principal's judgments and expected gain scores in which we were interested was stable enough for our purpose.

The two main questions the study attempted to answer, how valid principals' judgments are on the average and whether some principals judgments are more valid then others' will be answered by examining the distributions of principals' mean correlations and by analysis of variance. If there are significant differences in the validities of judgments made by different principals, we will ask what lies behind those differences.

The third question, which has to do with factors related to the size of the validity coefficients, was answered by a series of analyses of variance, one per principal. The set of correlations calculated for each principal was submitted to an analysis of variance in which the correlation between the principal's judgments and teachers' EGS's was the dependent variable. *Pupil Ability* (high or low), *Subject* (reading or mathematics), *Role* (I, II, or III), and, for those principals who recorded judgments of teachers in two or more grades, *Grade*, were the independent variables. The design was a four-way factorial;[3] with the difference between correlations based on different halves of the same class provided the estimate of error.

It should be noted that this error estimate reflected variations due to sampling of pupils from the population represented by the pupils in the same class only; it did not reflect variations due to sampling of teachers. The results

obtained are therefore not, strictly speaking, generalizable to other teachers but only to other pupils with these same teachers.

The purpose of the analysis was, of course, to examine the relationship between the dependent variable, the validity of the principal's judgment, and the independent variables as well as interactions between them. Since results for any one principal are of little interest, after estimating the components of the variance in each principal's correlation coefficients, we averaged the components across the sample to estimate the average importance of each factor in determining the magnitude of a correlation between any principal's judgments and EGS's of the same teachers.

Analysis of TPAI Data

Because of the small number of teachers rated on the TPAI in any one grade and school, it would not be possible to control grade, subject, and pupil ability by "blocking" them in the way we could in our study of the overall judgments. If and when the data become available, We will have to settle for a simple correlational analysis of the sample we obtain, one in which teachers and principals from different schools that use the same test are mixed together. Since in the certification process the instrument is used to compare teachers from different schools this may not be an inappropriate way to assess its validity and its relationship to principals' judgments.

NOTES

1. Medley, D.M. and Mitzel, H.E. Some behavioral correlates of teacher effectiveness, *Journal of Educational Psychology*. 1959, 50, 239-246.

2. Medley, D.M. *Teacher Competence and Teacher Effectiveness: A Review of Process-Product Research* Washington, D.C., American Association of Colleges for Teacher Education, 1977; Lara, A.V. *Pupil Ability as a Moderator of Correlations between Teacher Behavior Patterns and Pupil Gains in Reading and Mathematics.* Unpublished doctoral dissertation, Charlottesville, Virginia, University of Virginia, 1983.

3. For principals who rated only one grade, a three-way design was used.

RESULTS AND DISCUSSION

Distributions of Principals' Judgments

Before we examine the correlations between principals' judgments and the expected gains of their pupils, let us examine the kinds of judgments the principals record. Table 1 shows the distributions of the judgments of our sample of teachers recorded by the 46 principals on the three roles.

In these days when the public is convinced that there are so many incompetent teachers in the schools, these findings might make us wonder where they are. Fewer than 13% of these teachers were judged to be performing below average on any of the three roles. Indeed, according to their principals, these teachers were a remarkable group. About half of them were judged to be more effective than 85% of other teachers, and 13% were judged superior to *all* other teachers! This would be heartening news if we could believe it; but we can not. Like most people, when asked to rate or judge someone else, these principals are extremely lenient. Realizing how very difficult it is to make such judgments as these, and knowing the impact a low rating can have on a teacher's career, they hesitate to rate any but the most glaringly incompetent teachers very low.

Regardless of the validity or lack of validity of principals' judgments, the tendency that these figures clearly show for principals to overrate their teachers sharply limits the usefulness of their ratings as a basis for realistic decisions about teacher personnel. It also attenuates correlations between the judgments and other measures, including measures of teacher effectiveness.

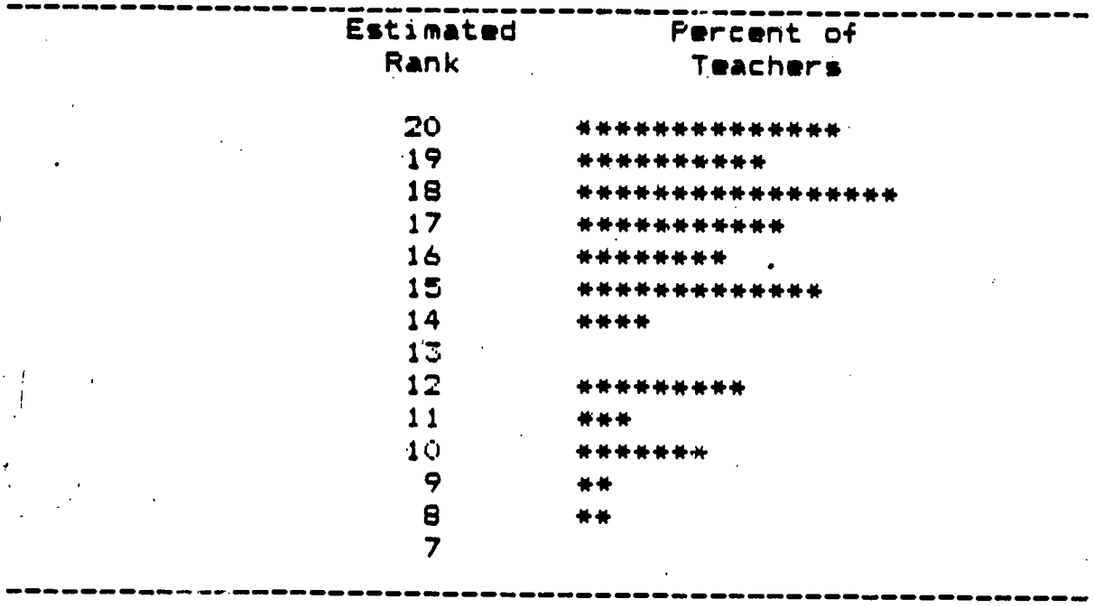
Note that the distributions for Roles II and III are identical. This does not mean, of course, that principals recorded identical judgments for each teacher; but it does mean that the amount of leniency displayed was the same on both roles.

In this study our interest centers primarily on judgments of teacher effectiveness in the first role, since it is the one which should relate most closely to EGS's (expected gain scores). Figure 3 shows the distribution of Role I judgments in graphic form. Note the crude modes at 20, 18, and 15. They suggest that the 20 levels of effectiveness used represented finer gradations than the principals felt comfortable in judging. Both of these

TABLE 1

DISTRIBUTION OF PRINCIPALS' JUDGMENTS OF EFFECTIVENESS OF TEACHERS IN PERFORMING THREE ROLES

Estimated Rank	Percent of Teachers		
	Role I Academic	Role II Affective	Role III Professional
20	13.7	13.3	13.3
19	9.5	8.4	8.4
18	16.7	18.3	18.3
17	11.0	9.1	9.1
16	8.0	6.1	6.1
15	12.9	12.9	12.9
14	3.8	4.2	4.2
13	0.4	2.3	2.3
12	8.7	6.5	6.5
11	2.7	1.1	1.1
10	7.2	10.3	10.3
9	1.9	1.9	1.9
8	1.5	2.3	2.3
7	0.4	1.1	1.1
6	0.0	0.8	0.8
5	0.4	0.8	0.8
4	0.0	0.0	0.0
3	0.4	0.4	0.4
2	0.4	0.0	0.0
1	0.4	0.4	0.4
Mean	15.6	15.4	15.4
S.D.	3.7	3.9	3.9



DISTRIBUTION OF PRINCIPALS' JUDGMENTS
OF TEACHER EFFECTIVENESS IN ROLE I

FIGURE 3

tendencies, the tendency to overrate teachers and the tendency not to use all available levels, tend to reduce the correlations between principals' judgments and EGS's.

A Sample Analysis

Before discussing how valid principals' ratings are it will be useful to present an example of the kind and amount of data generated for each principal. The complete set of validity coefficients calculated for one principal, Principal No. 70, is shown in Table 2. Principal No. 70 recorded judgments of four groups of teachers representing four different grades. The total number of coefficients calculated would therefore be 96.[1] Table 2 shows only the 48 whole-class values.

Note that the mean of all 48 correlations is .32; which means that the average validity of this principal's judgments is estimated to be .32. Since there is no reason to expect judgments on Roles II or III to correlate with EGS's, the mean Role I correlation, which is .40, is a better indicator of the validity of this principal's judgments then the overall mean of .32. Note also that Role I judgments made by this principal seem to be higher in grades 2 and 4, where they equal, respectively, .47 and .46, than they are in grades 3 and 6, where they are only .25 and .27.

Correlations based on samples as small as these, which contain only three or four teachers, are very unstable. But these are the sizes of the groups of teachers principals are called upon to compare; this is the evaluation task principals actually perform. A principal is likelier to need to decide which of three or four third grade teachers is the most competent than whether a third grade teacher is more competent than a sixth grade teacher.

There is considerable variation among this principal's correlations with EGS's in different subjects, grades, and levels of pupil ability. This variation was examined by means of an analysis of variance in a four-way factorial design as shown in Table 3.[2]

Notice that the only one of the factors studied that makes a statistically significant contribution to the validity of this principals' judgments is the interaction between grade taught and ability of pupil. From Table 2 we note that the differences between the validity coefficients for predicting gains of low-ability pupils and high-ability pupils for grades 2, 3, 4, and 6, respectively, were -.55, -.48, -.16, and +1.32.

TABLE 2

CORRELATIONS BETWEEN RATINGS OF TEACHERS AND PUPIL GAINS ACCORDING TO SUBJECT, GRADE, PUPIL ABILITY, AND TEACHER ROLE FOR PRINCIPAL NUMBER 70

Grade 2

Subject	Ability of Pupils	Teacher Role I	II	III	Average over Roles
Reading	Low	0.20	0.24	0.20	0.21
	High	0.64	0.85	0.64	0.71
	Average	0.42	0.54	0.42	0.46
Arithmetic	Low	0.19	0.20	0.19	0.20
	High	0.85	0.72	0.85	0.81
	Average	0.52	0.46	0.52	0.50
Averages over Subjects					
	Low	0.20	0.22	0.20	0.20
	High	0.75	0.78	0.75	0.76
Averages for Grade		0.47	0.48	0.47	0.47

Grade 3

Subject	Ability of Pupils	Teacher Role I	II	III	Average over Roles
Reading	Low	-0.13	-0.13	-0.07	-0.11
	High	0.66	0.66	0.54	0.62
	Average	0.26	0.26	0.23	0.25
Arithmetic	Low	0.15	0.15	0.15	0.15
	High	0.88	0.88	0.89	0.89
	Average	0.52	0.52	0.52	0.52
Averages over Subjects					
	Low	0.01	0.24	0.21	0.15
	High	0.49	0.49	0.72	0.56
Averages for Grade		0.25	0.45	0.46	0.39

TABLE 2 (Continued)

Grade 4

Subject	Ability of Pupils	Teacher Role I	II	III	Average over Roles
Reading	Low	0.83	0.73	-0.54	0.34
	High	0.14	0.10	-0.11	0.04
	Average	0.49	0.41	-0.32	0.19
Arithmetic		-0.07	0.35	0.28	0.19
	High	0.94	0.19	-0.97	0.05
	Average	0.43	0.27	-0.34	0.12
Averages over Subjects					
	Low	0.38	0.54	-0.13	0.26
	High	0.54	0.15	-0.54	0.05
Averages for Grade		0.46	0.34	-0.33	0.16

Grade 6

Subject	Ability of Pupils	Teacher Role I	II	III	Average over Roles
Reading	Low	0.90	0.90	0.90	0.90
	High	-0.01	-0.01	-0.01	-0.01
	Average	0.45	0.45	0.45	0.45
Arithmetic	Low	0.96	0.96	0.96	0.96
	High	-0.77	-0.77	-0.77	-0.77
	Average	0.10	0.10	0.10	0.10
Averages over Subjects					
	Low	0.93	0.93	0.93	0.93
	High	-0.39	-0.39	-0.39	-0.39
Averages for Grade		0.27	0.27	0.27	0.27

TABLE 2 (Continued)

AVERAGES OVER GRADES

Subject	Ability of Pupils	Teacher Role I	II	III	Average over Roles
Reading	Low	0.45	0.44	0.12	0.34
	High	0.36	0.40	0.26	0.34
	Average	0.40	0.42	0.19	0.34
Arithmetic	Low	0.31	0.42	0.40	0.37
	High	0.47	0.25	0.00	0.24
	Average	0.39	0.34	0.20	0.31
Averages over Subjects	Low	0.38	0.43	0.26	0.36
	High	0.42	0.33	0.13	0.29
Overall Averages		0.40	0.38	0.20	0.32

TABLE 3

ANALYSIS OF VARIANCE OF TEACHER RATINGS MADE BY PRINCIPAL NUMBER 70

Source of Variation	df	Sum of Squares	Mean Square	F
Role Rated	2	0.7854	0.393	1.073
Subject Tested	1	0.0213	0.021	0.058
Grade Taught	3	1.4227	0.474	1.296
Ability of Pupil	1	0.0988	0.099	0.270
Interaction R X S	2	0.0326	0.016	0.045
Interaction R X G	6	2.1678	0.361	0.988
Interaction R X A	2	0.1202	0.060	0.164
Interaction S X G	3	1.1716	0.391	1.068
Interaction S X A	1	0.1077	0.108	0.295
Interaction G X A	3	15.7792	5.260	14.377*
Interaction R X S X G	6	0.0308	0.005	0.014
Interaction R X S X A	2	0.6265	0.313	0.856
Interaction R X G X A	6	0.3061	0.051	0.139
Interaction S X G X A	3	0.9654	0.322	0.880
Interaction R X S X G X A	6	2.3265	0.388	1.060
Residual Variation	45	16.4630	0.366	
Total Variation	92	42.4256		

*P<.05

For some unknown reason, an apparent general tendency of this principal to prefer teachers who are more effective with bright pupils to teachers more effective with slow pupils seems to reverse itself rather dramatically in grade 6.

Finally, Table 4 shows the estimated proportions of the variance in a single validity coefficient that are associated with each of the 16 factors isolated in the analysis of variance shown in Table 3. More than half of the variance in this principal's correlations may be attributed to the interaction between grade and ability, and more than one-third to unexplained influences (residual variation). None of the other factors makes any appreciable contribution.

Factors in Validities of Principals' Judgments

These results might be of some interest to Principal No. 70 as descriptive of his performance with these teachers; but they are of little interest to anyone else because they lack generalizability. To obtain more useful results we performed an analysis of variance like this of the *6X12* correlations calculated for each principal in the sample (see Appendix A).

Proportions of variance associated with each of the 16 factors were averaged across all principals who recorded judgments on teachers in two or more grades. The results are shown in Table 5 for the 24 principals who recorded judgments of teachers in two or more grades. Table 6 shows the proportions for the 8 components of variance available in analyses of correlations for the 22 principals who recorded judgments of teachers in one grade only.

For comparison, the data for principals who recorded judgments of teachers in two or more grades on these eight factors are also shown in Table 6.

It is clear from Table 5 that grade level is the major factor related to the validity of principals' judgments of teachers. As a main effect, it accounts for more than one sixth of the variation; and the interaction Grade X Ability accounts for another tenth. In all, factors involving Grade account for more than 49% of the variations in validities of principals' judgments of teacher effectiveness; and identifiable factors not involving Grade for less than 18%. This should be compared with residual (unexplained) variance, which accounts for 33%.

TABLE 4

FACTORS IN RATINGS MADE BY PRINCIPAL NO. 70

Factor	Proportion of Variance
Role Rated	0.0045
Subject Tested	0
Grade Taught	0.0103
Ability of Pupil	0
Interaction R X S	0
Interaction R X G	0.0130
Interaction R X A	0
Interaction S X G	0.0117
Interaction S X A	0
Interaction G X A	0.5246
Interaction R X S X G	0
Interaction R X S X A	0.0054
Interaction R X G X A	0
Interaction S X G X A	0.0090
Interaction R X S X G X A	0.0686
Residual Variation	0.3529
TOTAL	1.0000

TABLE 5

FACTORS RELATED TO THE MAGNITUDES OF CORRELATIONS BETWEEN PRINCIPALS' RATINGS OF TEACHERS AND EXPECTED ACHIEVEMENT GAINS OF PUPILS IN THE TEACHERS' CLASSES
(Based on 24 Principals Who Rated Teachers in Two or More Grades)

FACTOR	PERCENT OF VARIANCE
Teacher Role Rated	1.6
Subject Tested	4.1
Grade Taught	17.7
Pupil Ability	4.5
Interaction R X S	0.1
Interaction R X G	3.5
Interaction R X A	0.3
Interaction S X G	7.8
Interaction S X A	7.1
Interaction G X A	10.3
Interaction R X S X G	0.5
Interaction R X S X A	0.0
Interaction R X G X A	0.4
Interaction S X G X A	8.2
Interaction R X S X G X A	0.9
Residual Variation	33.0
TOTAL	100.0

TABLE 6

FACTORS RELATED TO THE MAGNITUDES OF CORRELATIONS BETWEEN PRINCIPALS' RATINGS OF TEACHERS AND EXPECTED ACHIEVEMENT GAINS OF PUPILS IN THE TEACHERS' CLASSES

FACTOR	PERCENT OF VARIANCE	
	Teachers in One Grade Only Rated	Teachers in Two or More Grades Rated
Role Rated	6.4	3.3
Subject Tested	6.9	8.0
Pupil Ability	7.9	8.9
Interaction R X S	5.1	0.2
Interaction R X A	6.2	0.7
Interaction S X A	12.3	14.0
Interaction R X S X A	2.2	0.8
Residual Variation	53.0	64.1
TOTAL	100.0	100.0

Note that in the analyses in which Grades was not a factor all other factors combined account for less than half of the variation. The relationship of the ability level of the pupil whose expected gain score is correlated to the principal's judgment of teacher effectiveness, for example, is small.

More disturbing is the fact that the teaching role rated has virtually no relationship to the magnitude of the correlations, which suggests that judgments on Roles II and III must correlate with pupil gains just about as closely as Role I judgments. This is verified in Table 7, which shows the mean correlations by role and grade. The importance of grade level and the unimportance of role are both clearly apparent here.

Mean correlations seem to be high in odd-numbered grades (3 and 5), and low in even-numbered grades (2, 4, and 6). We have no ready explanation of this phenomenon; it may well be an artifact of the sample of principals.

Distributions of Validity Coefficients

Figure 4 shows the distribution of Role I validity coefficients (i.e., correlations between principals' judgments of teacher effectiveness in Role I and expected gains of the average pupil in a grade) across the sample of 87 grade groups rated. (The picture is much the same for judgments on Roles II and III, which are not shown.) The range is great, running (approximately) from -.75 to +.85; and the distribution does not depart much from normality.

The analysis of variance shown in Table 8 was designed to indicate whether this wide range is evidence that some principals' judgments are more valid than those of other principals, as the figure suggests. The F-ratio for differences between mean validities of different principals was only 1.19, which does not justify rejection of the hypothesis that there are no differences in the abilities of different principals to judge how effective a teacher is. This conclusion is based on the fact that judgments of teachers in different grades by the same principal vary almost as much as judgments made by different principals.

The F-ratio for interaction between principal and role of 1.39 is also small, so the hypotheses that it makes no difference which role is being rated cannot be rejected either.

TABLE 7

MEAN CORRELATIONS BETWEEN PRINCIPALS' RATINGS OF TEACHERS ON THREE ROLES AND EXPECTED GAINS OF PUPILS IN THE TEACHERS' CLASSES

GRADE	N	ROLE 1	ROLE 2	ROLE 3	AVERAGE
2	30	0.20	0.13	0.02	0.12
3	10	0.26	0.24	0.22	0.24
4	12	0.16	0.10	0.05	0.10
5	16	0.23	0.22	0.25	0.23
6	19	0.17	0.24	0.13	0.18
OVERALL	87	0.20	0.19	0.13	0.17

Correlation	Frequency
0.91 to 1.00	
0.81 to 0.90	*****
0.71 to 0.80	****
0.61 to 0.70	***
0.51 to 0.60	******
0.41 to 0.50	**********
0.31 to 0.40	**********
0.21 to 0.30	*********
0.11 to 0.20	******
0.01 to 0.10	******
-0.09 to 0.00	*****
-0.19 to -0.10	*****
-0.29 to -0.20	***
-0.39 to -0.30	*****
-0.49 to -0.40	*
-0.59 to -0.50	**
-0.69 to -0.60	*
-0.79 to -0.70	*
TOTAL	87

DISTRIBUTION OF CORRELATIONS BETWEEN PRINCIPALS' ROLE I RATINGS OF TEACHERS AND EXPECTED ACHIEVEMENT GAINS OF STUDENTS IN THE TEACHERS' CLASSES

FIGURE 4]

TABLE 8

ANALYSIS OF VARIANCE OF CORRELATIONS BETWEEN PRINCIPALS' RATINGS OF 87 GROUPS OF TEACHERS ON THREE ROLES AND EXPECTED ACHIEVEMENT GAINS OF STUDENTS IN THE TEACHERS' CLASSES

SOURCE OF VARIATION	D.F.	SUM OF SQUARES	MEAN SQUARE	F-RATIO
Role Rated	2	0.3488	0.174	0.48
Principal	45	19.4288	0.432	1.19
Group(same principal)	41	14.9144	0.364	10.05*
Interaction (RXP)	90	4.5356	0.050	1.39
Residual variation	82	2.9670	0.036	
TOTAL VARIATION	260	42.1946		

What is significant (F = 10.05) is the difference between validities of judgments of groups of teachers of different grades made by the same principal. Since each such group is made up of different individual teachers, the safest conclusion to draw is that it is easier to judge differences in effectiveness of some teachers than of others.

Concluding Observations

Despite our best efforts we have not been able to develop any credible evidence to indicate that principals' judgments of teacher effectiveness have any validity as predictors of how much pupils may be expected to learn about reading or arithmetic from them. The mean correlation between a principal's judgment of a teacher's effectiveness in teaching subject matter and expected achievement gains of the average pupil in that teacher's class in this study was only .20. A correlation of this size indicates that only four percent of the variance in principals' judgments reflects differences in teacher effectiveness; 96% of what these judgments indicate has nothing to do with teacher effectiveness.

These data do little to encourage us to believe that how valid a principal's judgment is depends on who the principal is, either. But here the small numbers of teachers in each group may be relevant. The range of estimated validities of different principals' judgments was very wide; but so was the range of estimates of validity of judgments made by the same principal.

When we studied the variations in estimates of the validity of judgments made by a single principal, we found that the major source of such variation was differences between groups of teachers of different grades. The parsimonious interpretation of this is that it is harder to judge the effectiveness of some teachers than others, and that this may be a function of grade taught.

Teachers are being evaluated all over the place by methods that are not detectably better than chance. If decisions about which teacher to certify, which to hire, which to award tenure to, and (soon) which deserve recognition as outstanding, were decided by a lottery they would be only a shade less accurate than the ones being made on the basis of principals' judgments of teacher effectiveness. It is time the profession accepted this disagreeable fact and did something about it.

NOTES

1. If a principal rated groups of teachers in *G* grades there would be *GX12* correlations, corresponding to *G* grades, 3 roles, 2 levels of pupil ability, 2 subjects, and 2 half-classes.

2. For readers interested in such matters, it should be noted that in three instances, separate estimates of the same correlation based on random halves of the same class were not available; hence 3 degrees of freedom for estimating residual variation were lost; the table therefore shows 45 degrees of freedom for residual variation (instead of 48) and 92 degrees of freedom for total variation (instead of 95).