DOCUMENT RESUME

ED 231 852 TM 830 381

AUTHOR Linn, Robert L.
TITLE Measuring School Effectiveness: How Achievement Data Can and Cannot Be Used.
PUB DATE Apr 83
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070) -- Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Tests; Criterion Referenced Tests; *Measurement Objectives; Norm Referenced Tests; Pretests Posttests; *Regression (Statistics); *School Effectiveness; *Scores; Socioeconomic Influences; Test Construction; Test Selection
IDENTIFIERS Survey Achievement Testing

ABSTRACT

In considering the problem of measuring achievement for the evaluation of school effectiveness, there are at least three questions that need to be answered: (1) What is to be measured? (2) How is it to be measured? (3) How are the results to be analyzed? Following a discussion related to the first two questions--determining content objectives and selecting or constructing tests that match the school's curriculum--attention is focused on the problems of translating test results into measures of school effectiveness. Primary consideration is given to what kinds of test scores should be used for analysis. The following types of scores are discussed: (1) global scores from survey tests, including the use of different forms of the same test; (2) average scores on a norm-referenced test or passing rates on a criterion-referenced test--including ranking in terms of status scores or trends in means for a grade, use of an SES indicator to adjust scores, and use of regression analysis to adjust for bias in mean gain scores; and (3) pretest/posttest scores, including three approaches for going beyond discussions of school means. The author concludes that comparisons of observed posttest results to those predicted from a regression of posttest on pretest scores seems the soundest approach to using achievement data as indices of school effectiveness. (LC)

Measuring School Effectiveness: How Achievement Data Can and Cannot be Used

Robert L. Linn

University of Illinois at Urbana-Champaign

Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, April 1983

Frechtling (1983) has identified a number of measurement issues and dilemmas in the evaluation of school effectiveness. She has divided these into two general areas: the measurement of achievement and the measurement of school processes and climate. The focus of this paper is on the issues in just one of these areas: the measurement of achievement. Although these comments will not provide anything like definitive answers to the tough problems that Frechtling has identified, they will highlight advantages and disadvantages of particular approaches, and show why some approaches are preferable to other commonly used ones.

In considering the problem of measuring achievement for the evaluation of school effectiveness, there are at least three questions that need to be answered: What is to be measured? How is it to be measured? and How are the results to be analyzed? Although the third question sounds more like a statistical question than a measurement question, it is an essential component of the validity of any inferences about the school effectiveness based upon student achievement data. Indeed, the flaws that Frechtling mentioned in regard to the four practices of using average scores, average gains, passing rates or differences in passing rates are largely the consequence of analytic shortcomings for the desired inference from the data.

What and How to Measure

The questions of what and how to measure is obviously of central importance. A widely circulated ETS phamplet entitled Selecting an Achievement Test: Principles and Procedures (ETS, 1969) astutely notes that "Before deciding what we want to test . . . , we must have a clear identification of what we want to teach" (p. 14). What are the curricular objectives and in terms of which of these objectives should school

effectiveness be evaluated? After these questions have been answered the process of constructing or selecting appropriate tests can begin. Too frequently, this first step of deciding what should be measured is given too little attention, indeed the process may even be reversed, that is, a test may already be in place and its degree of match to the curriculum judged after the fact.

Rowan, Bossert and Dwyer (1983, p. 25) have noted that "Past research has defined school effectiveness narrowly as instructional effectiveness and has measured this construct using standardized achievement tests." They go on to argue that this narrow approach ignores many important goals. This is quite true, however, instructional goals are clearly important and, as Rowan, et al. also note, there are substantial difficulties in adequately measuring even this aspect of effectiveness.

A difficult issue in defining the knowledge and skills to be tested is the question of whether measurement should be limited to a core that is common to all schools to be studied or include relatively unique objectives that are pursued by only a few schools. Limitation to a common core may conceal some of the most important differences between schools. On the other hand, including items that measure objectives unique to a few schools may greatly increase the testing burden and be considered unfair to schools that do not pursue those objectives.

To the extent that it is feasible, it is important to go beyond the common core and provide some coverage of content that is emphasized by some but not other schools. The two categories of items must be treated separately in the analysis and doing so may complicate conclusions. Consider, for example, two hypothetical schools: the curriculum at school A includes objectives in computer literacy while the school B curriculum does

not. Skills in basic arithmetic operations, on the other hand, are part of the curriculum at both schools. The analysis of the achievement results indicate that school B is more effective than school A in terms of arithmetic operations, but school A is more effective than B in terms of the computer literacy measure. There is no simple answer to the question of which school is more effective That judgment depends on the value attached to the two areas of achievement and that judgment will surely vary from one individual to another. But the greater complexity is surely a more complete picture than would be obtained from a comparison limited to the common core of arithmetic operations.

Once the content objectives have been determined the process of test selection or construction can begin. At this point there is apt to be a debate about the relative merits of norm-referenced and criterion-referenced measures. However, these labels should not be the primary consideration. It is an analysis of test content in which judgments are made about the match between the test and the curriculum and the likely sensitivity of the test to school differences that is crucial.

Test publishers rely on fairly similar techniques that depend on careful analysis of widely used curriculum materials to define the content coverage of their tests. They produce tests with similar names. The test scores of the most nearly comparable tests of different publishers are highly correlated (e.f., Bianchini & Loret, 1974). These similarities conceal potentially important differences in detailed content coverage and the match of coverage to the curriculum, however. Detailed comparative analyses such as those reported by Hoepfner (1978) for reading tests and by Porter, Schmidt, Floden and Freeman (1978) for mathematics tests reveal surprisingly large differences between tests. Futhermore,

the degree of overlap between what is taught and what is tested has been found to be closely related to performance (e.g., Bianchini, 1978; Leinhart, 1983).

The importance of overlap between test content and either curriculum materials or teacher reports of instruction is illustrated by Leinhart's (1983) summary of the results of two studies. In one of those studies teachers identified curricula used with each student. A computer list of the words in the curriculum materials used for each student. A list of words on the test was also compiled. These lists were then used to obtain estimates of overlap for each student. With pretest partialled out, the correlation between the posttest and overlap was .38. Similar results were obtained using instruction-based estimates of overlap. Such results suggest that consideration of overlap may be critical in studies of school effectiveness.

Analysis

Although the choice of the tests to be used in an evaluation of school effectiveness is of crucial importance, no further consideration will be given to that issue here. It is a topic that is given considerable attention, not only in most tests and measurement textbooks, but in most test manuals. This is not to say, of course, that the advice in these sources is always heeded. But even if the tests are chosen with care, several obstacles stand in the way of translating the test results into measures of school effectiveness. Some of these have been clearly stated by Frechtling (1983): should status or gain be analyzed and should averages or criterion attainment be used? It seems to me, however, that a prior question is what scores should be used?

A single global score in mathematics may serve some purposes and may be the only mathematics score with sufficient reliability at the individual student level. However, the focus in a study of school effectiveness is at a different level and global scores may conceal differences that exist for finer breakdowns of the content. The intermediate level mathematics survey test of the Metropolitan Achievement Tests (Prescott, Balow, Hogan & Farr, 1978), for example, consists of 50 items that span 7 content areas. These are numeration, geometry and measurement, problem solving, whole number operations, laws and properties of operations, fraction and decimal operations, and graphs and statistics. The number of items per content area ranges from 3 to 13. Although the number of items in a single content strand is too few for reliable individual measurement, separate content scores for the content strands may have utility at the school level. Of course, more items per strand would yield greater fidelity and this could be accomplished by using the Intermediate MAT Mathematics Instructional Tests which cover the same 7 content areas with between 18 and 42 items per area for a total of 204 items. The tradeoff of greater fidelity for these areas of mathematics is apt to be a narrower bandwidth, i.e., better measurement in mathematics at the expense of less coverage in other areas.

An alternative approach that can enhance both fidelity and bandwidth at the school level is have students respond to different tests. For example, the number of items in the content strand would be doubled by administering Form JS of the MAT Survey Battery to half the students and Form KS to the other half, thereby providing between 6 and 26 items per content strand for school level analysis and still having complete survey test scores for individual students.

One or Two Administrations. Comparing schools in the results of a single test administration whether in terms of average scores on a norm-referenced test or passing rates on a criterion-referenced test has serious flaws. As Frechtling (1983, p. 3) noted, one is apt to find that "the apparently most successful school is that serving the wealthiest students from the best educated families." Even when trends in test scores, e.g. the mean scores for a particular grade of a school over several years (Phi Delta Kappan, 1980) are used, the increases or decreases are apt to reflect changes in "the socioeconomic composition of a school's student body." (Rowan, Bossert, & Dywer, 1983; see also, Rowan & Denk, 1982). Hence simple ranking in terms of status scores or trends in means for a grade cannot be considered a fair measure of school effectiveness. At a minimum comparisons must take into account differences in socio-economic status. Some test publishers provide special report services that incorporate adjustments for differences in socio-economic status. The MAT again peovides an illustration of this approach. The SES predicted achievement report (The Psychological Corporation, 1981) provides comparison of obtained mean achievement scores for a school to ranges of scores predicted from a parental education index. Schools are located in one of five score bands based on a regression analysis of school means in the MAT school norms. Nationally, schools would be expected to be distributed with about 10%, 20%, 40%, 20% and 10% of the schools in bands 1 through 5 respectively.

The use of an SES indicator to adjust scores is an improvement over simple ranking on achievement scores, but the adjustment may not be adequate. In Cronbach's (1982, p. 191) terminology, SES is almost surely an "incomplete covariate." That is, it provides only a partial adjustment

for preexisting differences outside the control of the school. Finding the complete covariate is undoubtedly an illusary goal but prior achievement probably provides the closest feasible approximation. For this reason, estimates of school effectiveness are more dependable when pretest results are used to adjust posttest performance. Gain scores are the simplest, but not the best way of making the desired adjustment. A regression effect may bias the results of an analysis of mean gain scores. While slightly more complicated, a regression analysis, alleviates this problem.

A regression approach to deriving indices of school effectiveness has been described by Dyer (1966, 1970a, 1970b) and several authors (e.g. Dyer, Linn & Patton, 1960; Forsyth, 1973; Marco, 1974; Marco, Murphy & Quirk, 1976; Rowan & Denk, 1982) have investigated variations and properties of this general approach. In its simplest form, school posttest means are regressed on school pretest means and school performance indices are based on deviations of observed posttest means from their predicted values. Other predictors, e.g. means on other pretests or measures of SES, may also be incorporated in the regression, but are apt to improve the predictive power of the pretest relatively little.

A dilemma in this approach, which also applies to the use of average gains, is caused by missing data. Some students will have pretest scores but no posttest scores while the converse is true of others. As shown by Dyer, Linn and Patton (1969) the results for cases with complete data (both pretest and posttest), may differ from those based on means for all students with one or both scores. The complete data results may be based on only a small fraction of the students served by a school where mobility is high. On the other hand, it seems unreasonable to attribute effects to a school based on changes in the student body due to mobility. Hence,

the use of only cases with both pretest and posttest scores seems preferable. Student mobility is a relevant variable to consider in interpreting those results, however.

In her discussion of the use of average gain scores, Frechtling (1983) observed that there seems to be little consistency between school gains from one year to the next, with correlations between mean gains of only about .2 or .3. She concluded that "either school effectiveness is a very fragile thing or the metric used, the gain score, has serious problems. As was previously stated, gain scores less adequate than scores based on a regression approach. However, the evidence suggests that the latter approach may produce results that are no less fragile. Forsyth (1973) investigated the stability of school residuals from regressions of school mean posttest on pretest scores from one year to the next. The median correlation between residuals for 10 different scores was only .28, with a range from a low of .11 for Ability to do Quantitative Thinking to a high of .50 for Social Studies. Similar results have been reported by Jencks, et al. (1972) and by Rowan and Denk (1982). Although these results may be more fragile than seems desirable, they may also reflect reality, at least at the level of general composite scores. It may be that somewhat more stable scores would be obtained by more specific content scores of the type illustrated earlier. It also may be that scores for content areas where there is less uniform agreement on coverage across schools, such as the computer literacy example used above, would yield more stable results from year to year than those obtained for the core areas. Furthermore, even this limited amount of stability may be sufficient for contrasting extremes, e.g. the 10% of the schools with the largest positive residuals with the 10% that have the largest negative residuals.

Beyond Averages. So far the focus has been only on school means. This focus is understandable, but it ignores other possible differences between schools that are of potential interest. Two schools may appear quite similar in terms of average scores but differ considerably in the performance of the highest and lowest achieving schools. That is, a given average gain may be obtained as the result of large gains for initially, low scoring students and only modest gains for initially low scoring students. Conversely, the same average gain may be achieved by large gains at the upper end of the distribution at the expense of gains at the lower end.

Two approaches have been suggested for going beyond school means. Dyer, Linn and Patton (1969) used the 20th and 80th within-school percentiles in addition to school means. Posttest scores at the 80th percentile, for example, were regressed on pretest scores at the 80th percentiles. A comparable analysis was performed using within-school 20th percentile scores. Schools that are identified as effective using means were not necessarily the same as those that were identified using 20th or 80th percentile points.

An alternative approach that takes into account differential effects on initially high and low scoring students within a school has been proposed by Burstein, Linn and Capell (1978). In the latter approach, within-school regressions of student posttest on pretest scores are computed. The within-school slopes are used along with results based on school means to describe school performance. Attempts are then made to explain both sets of results in terms of school process variables.

The gain in percent passing a prespecified standard on a criterion-referenced test, which was described by Frechtling (1983) and is the basis of the analysis reported by Clark and McCarthy (1983), may also

give information not contained in an analysis of average gains. However, this approach is less satisfactory than either of the two approaches just described. The choice of a standard is fraught with difficulty and the percentage increase metric has undesirable properties. It seems unreasonable, for example, to consider a change from 10 to 20 percent passing comparable to a change from 85 to 95% passing.

Conclusion

While no panacea, comparisons of observed posttest results to that predicted from a regression of posttest on pretest seems the soundest approach. Within-school points in the distribution (e.g., 20th and 80th percentile) or within-school regressions as well as means are potentially relevant. The scores should be as content specific as feasible and range over both common and relatively unqiue objectives. Even so, the evidence suggests that only a modest degree of stability can be expected from one year to the next. Hence, it seems wise to avoid drawing conclusions from small differences in indices of effectiveness.

REFERENCES

Bianchini, J. C. Achievement tests and differential norms. In M. J. Wargo & D. R. Green (Eds.) Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey, CA: CTB/McGraw-Hill, 1978.

Bianchini, J. C. & Loret, P. G. Anchor test study final report. Project report and volumes 1 through 30, and the anchor test study volumes 31 through 33, 1974. (ERIC Documentation Reproduction Services Numbers ED092601 through ED092634.)

Burstein, L., Linn, R. L., & Capell, F. J. Analyzing multilevel data in the presence of heterogeneous within-class regressions. Journal of Educational Statistics, 1978, 3, 347-383.

Clark, T. A. & McCarthy, D. P. School improvement in New York City: The evolution of a project. Educational Researcher, 1983, 12, No. 4, 17-24.

Cronbach, L. J. Designing evaluations of educational and social programs, San Francisco: Jossey-Bass, 1982.

Dyer, H. S. The Pennsylvania study. Science Foundation, 1966, 50, 242-248.

Dyer, H. S. Toward objective criteria of professional accountability in the schools of New York City. Phi Delta Kappan, 1970a, 52, 206-211.

Dyer, H. S. Can we measure the performance of educational systems? National Association of Secondary School Principals' Bulletin, 1970b, 54, 96-105.

Dyer, H. S., Linn, R. L. & Patton, M. J. A comparison of four methods of obtaining observed and predicted school system means on achievement tests. American Educational Research Journal, 1969, 6, 591-605.

Educational Testing Service, Selecting an Achievement test: principles and procedures. Princeton, NJ: Educational Testing Service, 1969.

Forsyth, R. A. Some empirical results related to the stability of performance indicators in Dyer's student change model of an educational system. Journal of Educational Rsearch, 1973, 10, 7-12.

Frechtling, J. A. Problems in evaluating school effectiveness: A shopping list of dilemmas. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, April, 1983.

Hoepfner, R. Achievement test selection for program evaluation. In M. J. Wargo & D. R. Green (Eds.), Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey, CA: CTB/McGraw-Hill, 1978.

Jencks, C. L., Smith, M., Acland, H., Bane, M. J., Cohen, D. K., Gintis, H., Heyns, B. L. & Michaelson, S. Inequality: A reassessment of the effects of family and schooling in American. New York: Basic Books, 1972.

Leinhart, G. Overlap: Testing whether it is taught. In G. F. Madaus (Ed.), The courts, validity, and minimum competency testing. Boston: Kluever-Nijhoff, 1983.

Marco, G. L. A comparison of selected school effectiveness measures based on longitudinal data. Journal of Educational Measurement, 1974, 11, 225-234.

Marco, G. L., Murphy, R. T. & Quirk, T. J. A classification scheme for methods of using student data to assess school effectiveness. Journal of Educational Measurement, 1976, 13, 243-252.

Phi Delta Kappan. Why do some urban schools succeed? The Phi Delta Kappan study of exceptional urban elementary schools. Bloomington, Indiana: Phi Delta Kappan and Indiana University, 1980.

Porter, A. C., Schmidt, W. H., Floden, R. E. & Freeman, P. J. Impact on what?: The importance of content covered. East Lansing, MI: Institute for Research on Teaching, Michigan State University, 1978.

Prescott, G. A., Balow, I. H., Hogan, T. P, & Farr, R. C. Metropolitan Achievement Tests, teacher's manual for administering and interpreting. New York: The Psychological Corporation, 1978.

The Psychological Corporation, The SES predicted achievement report: Its purpose, development and use. Metropolitan Achievement Tests, Special Report No. 25. New York, The Psychological Corporation, 1981.

Rowan, B. & Denk, C. E. Modeling the academic performance of schools using longitudinal data: An analysis of school effectiveness measures and school and principal effects on school-level achievement. San Francisco, California: Far West Laboratory for Educational Research and Development, 1982.

Rowan, B., Bossert, S. T. & Dwyer, D. C. Research on effective schools: A cautionary note. Educational Researcher, 1983, 12, No. 4, 24-31.