Journal of Teaching and Learning

Volume 2 | Issue 1 Article 2

4-1-1976

Education, Evaluation, and the Metrics of Learning

Barbara Heyns

Education, Evaluation, and the Metrics of Learning*

Barbara Heyns
University of California at Berkeley

The most common critique of recent large-scale cross-sectional research is that the survey design is not longitudinal, and that the critical dependent variable, relative achievement level, is not an accurate assessment of learning. Studies which have included longitudinal test data, however, have rarely demonstrated significant effects; such studies are then criticized because they involve ex post facto designs with poorly matched control groups and analysis of covariance adjustment techniques (Lord, 1967; Lord, 1969; Campbell and Erlebacher, 1970; Cronbach and Furby, 1970; Rossi and Williams, 1972). Randomization is frequently an impossibility in quasi-experimental research designs, and as Lord has argued, " . . . there is no logical or statistical procedure . . . which . . . makes proper allowances for uncontrolled preexisting differences between groups" (Lord, 1967, p. 305). The present paper is oriented toward a more fundamental critique of the use of test scores in a longitudinal analysis, and a partial explanation of the inability to demonstrate significant differential learning in either natural or experimental settings. I shall argue that what is needed to evaluate the impact of programs on the academic performance of children is an empirically verified measure of learning, rather than changes in relative position on standardized tests. At the outset, I will discuss common concepts of test theory and measurement, and provide examples

*Paper prepared for the American Sociological Association meetings, Montreal, 1974.

Professor Heyns is on the faculty of the University of California at Berkeley. Research support for this paper is from the Office of Economic Opportunity, Grant No. 10301. Permission to cite or quote should be requested from the author.

of how differences in metric change the interpretations of research findings. The purpose is twofold: to clarify the assumptions in using existing test metrics, and to suggest ways of constructing more valid indicators of change.

The psychometric approach to measuring differential ability or achievement is often cited as "the most important technical contribution psychology has made to the guidance of human affairs" (Cronbach, 1970, p. 197). In broad outline, the methods and assumptions employed are similar to a variety of forms of personality assessment prevalent in psychology. The model is adopted directly from the physical sciences, and textbooks abound with analogies between measuring heat and measuring intelligence. The model essentially construes ability (or achievement, anxiety or self-esteem) as a trait, or a construct which is present by degree in individuals. An adage of the psychometrician is that if a thing exists, it exists in certain amounts, and therefore can be measured. The construct is assumed to inhere in individuals, as an innate trait, behavioral predisposition, or as a fluid property. Constructs are related to behavior through "semantic" (Lord and Novick, 1968) or "epistemic" definitions (Torgerson, 1958), which form the rules of correspondence between theoretical constructs and the domain of observable behavior. A variable is considered an indicator of the construct if the expected value varies systematically with respect to the construct. A variable is considered a measure if, and only if, the expected value of the indicator increases monotonically with respect to the construct. Achievement test scores are therefore valid measures of the latent, unobserved construct achievement, if they are isomorphic and increase monotonically with respect to the construct.

Viewed in this manner, the problem of measurement is selecting the set of items which provide a mapping from test scores to the construct achievement. This generally reduces to defining the set of items which best differentiate between persons with more or less of the postulated construct. Achievement test items are selected which best discriminate between high and low scoring pupils, for example. Such a process is justified as a means of refining the measure, although it also has the effect of selecting the subtest of items most highly related to each other. Achievement tests are not generally scrutinized as closely nor scaled as adequately as were initial tests of intelligence; in fact, most achievement tests are routinely compared to I.Q. tests in order to validate the postulated ordinal relationships and in order to impute an interval scale and a known distribution. The degree to which a construct exists or is meaningful is not questioned if the resulting measure can be shown to be predictive of differential performance. The problem of defining a mapping function which relates the measure to the construct is common to all tests of cognitive ability. It has led to assertions such as that of Edwin Boring, that "intelligence is what intelligence tests measure." The parallel assertion would be equally valid, although perhaps less satisfying, that reading is what tests of reading comprehension measure.

The levels of measurement involved in test construction are important for an understanding of the logic of establishing numerical relationships and assessing rates of growth. A level of measurement is distinguished by the relative complexity of the mathematical system with which it is associated; levels admit certain kinds of transformations and operations which may be performed meaningfully within the system (cf. Guilford, 1954; Torgerson, 1958; Kerlinger, 1964; Lord and Novick, 1968; Krantz, Luce, and Supps, 1971). Nominal measurement is used only for purposes of classifying objects into mutually exclusive subsets. In order to do so, one must be able to apply the identity and equality functions:

$$(a = b) \text{ or } (a \neq b), \text{ but not both} \qquad [\text{Eq. } 1]$$

$$[(a = b) \text{ and } (b = c)], \text{ then } (a = c) \qquad [\text{Eq. } 2]$$

A nominal category is not strictly measurement, since a numerical label can be attached only for the purpose of classification, without any assumptions of order or magnitude.

The second or ordinal level of measurement presupposes nominal properties, and in addition requires that the transivity postulate:

$$\text{If } [(a > b) \text{ and } (b > c)], \text{ then } (a > c). \quad [\text{Eq. } 3]$$

The capacity to order subsets meaningfully is thus permitted, and any scale which yields the same relative order is an admissable transformation.

The level of measurement most commonly assumed in achievement tests requires interval measurement. An interval scale specifies a direct mapping between behavioral elements and the real numbers; the zero point cannot be set and the metric or interval length is arbitrary. Interval scales can be subjected to any linear transformation, such as addition or subtraction. We may say for example that the difference between students scoring six and eight is equal to the difference between students scoring two and four; however, we cannot assert that the achievement of students scoring eight is twice as great as those scoring four.

A fourth level, that of ratio measurement, allows for multiplicative relationships, since the existence of a zero point is fixed. A ratio scale permits one to argue that one unit has half as much or three times more than the amount of achievement as another. Ratio measurement is never claimed for achievement tests, although the literature abounds with conclusions which, strictly speaking, require it.

Ordinal measurement is a necessity for assessing differential ability or achievement. The assumption is made that ability is cumulative, and that it is possible to select a subset of items which can be ordered in such a manner that a correct response on the $n+1^{st}$ item presupposes correct responses on the n preceding items. Such a relationship is basic for asserting ordinal properties to a measure and for the construction of scales. Interval measurement is imputed arbitrarily to tests of ability, based on pragmatic considerations. The most frequent rationale is distributional, and most tests are scaled to yield a normal distribution. Abelson and Tukey (1959) have suggested choosing scales so as to provide additivity of effects; however, most testing specifies a convenient distributional form which allows interval properties based on a normal curve. The distribution is justified by the central-limit theorem and by the empirical fit of measures. A substantial amount of work has been done in scaling items for intelligence tests and generating a subset of items which would resemble a Gutman scale, and also would be predictive of school achievement, defined by teacher's ratings or student performance on related criteria. For most achievement batteries, comparable work has not been done; instead, a substantial and linear relationship with intellectual capacity defined by I.Q. tests is taken as a sufficient rationale for imputing interval properties.

For any particular ability test, the assumptions present do not seem particularly onerous, although one might wish the constraints imposed were given more than passing mention. The difficulty in an analysis of data in which one is concerned with differential learning, not just relative position at one point in time, however, is large.

In order to assert that a gain based on a particular metric is isomorphic to learning, one must assume that learning is a monotonically increasing function described by changes in test score data, based on a particular metric. For achievement tests, the most commonly used metrics are raw scores, standardized raw scores, or grade equivalent scores. The measures are not typically linear transformations of each other and yield quite different learning profiles for students at different positions on the scale. For example, if student A scores 50 and student B scores 40 on a 60-item achievement pretest, the question of whether A's achievement is higher than B's is one of the validity and reliability of the test. Classical test theory defines an ordering in terms of true scores, and imputes a metric assuming a normal distribution of the construct achievement. Imagine, however, that students A and B scored 55 and 50 respectively on a parallel form posttest. How would one determine which student learned the most? The interval [50, 55] is less than the interval [40, 50]; on a test in which all items were equally difficult, the intuitive response that student B had learned the most would be correct. However, transforming scores to standardized values, normed on a national sample, the gains appear equal, at about 1.4 standard deviations. In grade equivalent units, student A gained 1.3 years, while student B gained only .8 of a year. Such disparities are not unusual in comparing achievement metrics. The transformations which are typically used to evaluate growth fundamentally depend on assumptions about the metrics of achievement. Particularly for Sociologists, who are often interested in comparing the relative growth of students from diverse socioeconomic backgrounds, interpretations of relative growth depend on the metrics of learning and the position on the scale originally.

A more precise example of the difficulties inherent in comparing test results and inferring learning is given in Table 1. The data is based on the Metropolitan Achievement Test, Word Knowledge, with gains computed as the difference between a fall 1971 pretest and a spring 1972 posttest on a parallel form of the intermediate battery. The sample is a group of white sixth grade students in Atlanta, presented by levels of mother's education. Raw scores reflect the absolute number of correct responses on the test of student vocabulary. The average white student in Atlanta gained five and one-half words during the sixth grade; however, the gains were inversely related to mother's education. Grade equivalent gains yield precisely the opposite conclusion with respect to background, while either standardized metric yields quite ambiguous results.

The explanation of the observed pattern does not depend on threshold effects, or on non-linearity between a pretest and posttest. It depends on the data transformations involved in changing metrics. Raw scores are not linearly related to either normalized

TABLE 1

COMPARISON OF GAIN SCORES BY METRIC, METROPOLITAN ACHIEVEMENT TEST, WORD KNOWLEDGE FOR WHITE SIXTH GRADE STUDENTS IN ATLANTA, BY MOTHER'S EDUCATION, FALL 1971 - SPRING 1972

Mother's Education	Raw Score	Standard Score[1]	Grade Equivalent[1]	Standardized Scores, Atlanta Only[2]
White	5.5	4.0	.85	.28
0-11 years	6.6	4.0	.53	.55
12 years	6.3	4.0	1.02	.81
13+ years	5.0	4.1	1.19	-1.04

[1] Published norms, based on national sample.

[2] Standardized raw scores, for Atlanta population. The total gain is not equal to 0 because non-white students were included in the sample. Scores were normalized prior to standardization.

standard scores or to grade equivalents, although they are used to create both. Raw scores plotted against normalized standard scores on the Metropolitan Achievement Tests yields an S-shaped distribution, largely because the actual scores are slightly more peaked than a normal distribution. For consecutive batteries, standardized scores are quite erratic and lead one to conclude that gains are not linear across grades.

Grade equivalent gain scores are computed by interpolating raw scores between pupils of different ages, and imputing a score based on the average attained at a particular grade level. Such scores are relatively crude, have quite large standard errors, and are not advisable for assessing a particular individual's position. For the purpose of evaluation, however, they embody a metric which is at least tied to the actual expected scores of pupils over time. Grade equivalents are not linearly related to either raw scores or standard scores, but give greater increments of gain as one increases in actual raw score. This implies that the gains are larger per raw score point for scores above the median, while the lower half of the distribution requires a greater actual improvement in numbers of questions correct to product an equivalent gain. For this reason, one finds in large cross-sectional surveys that the gap between white and black students in standardized scores can remain constant, at one standard deviation, while the gap in grade equivalence increases consistently (Coleman and Karweit, 1970).

In terms of learning, the various metrics incorporate different analytic assumptions. To utilize raw scores, one must assume the test items are scaled to produce equal intervals, which is tantamount to assuming items are equally difficult between and within forms. This is clearly not the case. Standardized scores transform all raw scores to a common distributional form based on a normal distribution. Standardized scores ignore any learning which does not alter the relative position of students, by equalizing the variances across time. Standardized scores are enormously influenced by the norming population and, as Table 1 demonstrates, yield quite different results when comparing subgroups to the nation.

Both standardized scores and grade equivalents embody the critical assumption that the total number of items correct is an ordinal scale that unambiguously orders individuals by relative achievement. All transformations of raw scores are designed to impute interval measurement; although this is a worthwile endeavor, it would make considerably more sense to examine the items and to determine whether students at some particular relative position or grade level position could correctly respond to the particular item, and then to weight the respective items by relative difficulty defined by the probability of a correct response. Such procedures would not assume that the total score is designed to order students according to some trait, nor necessarily assume that achievement is a construct.

The construct paradigm has become so much a part of the general theoretical orientation in education that it is difficult to imagine an alternative approach. For the sake of comparison, one might conceive of intelligence or achievement as a set of cognitive and behavioral skills. Rather than being an entity possessed in varying amounts, the skills would be divisible in a variety of ways. The position of individuals would be defined by the proportion of total skills known, and in relation to some universe of knowledge, not merely in relation to a set of peers. With such a notion of achievement, a sampling paradigm would be a much more fruitful perspective for developing measures of differential performance. Considerably more attention would be paid to how representative particular items were of the universe of skills a person had at his command, rather than just relative position. For example, if one wished to assess the vocabulary skills of a sixth grade child, it would be useful to know how closely correct responses reflected total words known. A finite universe of words exists which could be categorized by relative difficulty, for example. Items would be randomly selected to be representative of the categories, and correct responses weighted by the sampling probabilities. The major advantage of such a procedure would be that a test could be assumed to measure the size of total vocabulary, and the rates of learning could be assessed. Logical research questions to be posed would be how best to categorize the skills in question, and how to order subsets by complexity, difficulty, or relevance. Learning in this context would be measured as an increase in the proportion of skills the person could command, rather than just shifts in relative position. The metric underlying differential performance could be related to the abscLute level of information or skills, with errors due to sampling. In practice, the weighted scheme employed might be no more complex than the present indices computed to determine the relative difficulty of items. The advantage, however, would be a considerably less ambiguous metric for determining cognitive development. At present, the only conceptual paradigm which is available cannot be used to determine how much learning occurs, or what a differential rate of learning would be, without implausible assumptions. Such issues are crucial for the analysis of longitudinal data, irrespective of the study design.

In an effort to demonstrate the measurement variability introduced by assuming items are equally difficult, the items on the Metropolitan Achievement Test, Word Knowledge, administered in Atlanta in the fall of 1972 were analyzed. Each item was weighted by the proportion of students correctly responding to the question, and the resulting scores standardized. Table 2 presents the mean scores from the standardized raw scores and the standardized weighted raw scores, for Atlanta students, by particular number of correct responses. As is evident from the table, the variances in standardized weighted scores is quite considerable, and related to the position of students. This implies that a similar raw score could lead to very different actual positions when relative difficulty is controlled. The cluster of items which a student correctly answers could be more or less difficult than the raw score would indicate, and the error introduced could be as much as five or ten actual correct items. The degree of variability is as large as the average increment during one year of schooling, the increment which we typically call learning. Since both the weighted and unweighted scores would have the same reliability, the variance introduced is due to differential difficulty rather than to error. Quite obviously, one would want to extend the item analysis to see the degree to which difficulty influences estimates of learning; however, the necessary pretest scores by items was not available.

Although the present analysis has largely focused on practical difficulties, the theoretical issues are of no mean importance. Psychometricians have expressed concern with the long-standing division between learning theory and classical test theory (Atkinson and

TABLE 2

COMPARISON OF METROPOLITAN ACHIEVEMENT TEST STANDARDIZED SCORES AND SCORES WEIGHTED BY DIFFICULTY, ATLANTA 1972

Raw Scores	Mean Standardized Score	Weighted Standardized Score Mean	Weighted Standardized Score Variance
25	44	43.2	7.1
30	50	51.1	8.2
35	54	57.3	9.3
40	59	68.2	10.1
45	66	74.3	12.8
50	80	88.1	16.7

Paulson, 1971; Cotton and Harris, 1973). Achievement scores are generally considered less heritable than intelligence tests (Jensen, 1967), although this may reflect merely greater unreliability. The notions implicit in classical test theory lead one to attribute much change in relative position to unreliability; efforts to separate stability and unreliability have been made, although they assume either a constant variance or constant reliabilities over time (Heise, 1971; Wiley and Wiley, 1972; Armor, 1973). The general problem leads to what Bereiter has called the "unreliability-invalidity dilemma"; that is, the more reliable a particular measure, the higher is the test-retest correlation, and the less it varies under any experimental conditions. The lower the reliability, the greater the probability that the test is not measuring the same thing, and therefore the validity is low. Classical test theory assumes reliability is a property of the instrument, and is random with respect to particular individuals; yet it can easily be demonstrated that reliabilities, whether KR-20 as a measure of internal consistency or test-retest correlations, differ for different socioeconomic groups.

The argument offered herein is that we have at present no clear conceptual or empirical tools which lead to unambiguous assessments of learning. Changes in raw scores are not an ordinal scale of learning, unless one wishes to assume that different groups learn at different rates depending on the metric used. Available tests invite highly consistent results for evaluation, and that is that no changes in relative position occur which cannot be attributed to unreliability. What is needed is a more precise, empirically verified theory of learning with which to evaluate the progress of children.

Bibliography

Abelson, R. and Tukey, J. W. "Efficient conversion of non-metric information into metric information." American Statistical Association Proceedings of the Social Statistics Section, 1959, Pp. 226-230.

Armor, D. J. "Toward a unified theory of reliability for social measurement." The Rand Corporation, P-5264, 1974.

Atkinson, R. C. and Paulson, J. A. "An approach to the psychology of instruction." Psychological Bulletin 1972, 78 (1), Pp. 49-61.

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

Campbell, D. T. and Erlebacher, A. "How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful." In J. Hellmuth (Ed.), Compensatory Education: A national debate. Vol. III. The disadvantaged child. New York: Brunner/Mazel, 1970. Pp. 185-210.

Campbell, D. T. and Stanley, J. C. "Experimental and quasi-experimental designs for research on teaching." In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. Pp. 171-246.

Coleman, J. S. and Kurweit, N. L. Information systems and performance measures in schools. Educational Technology Publications, 1972.

Cotton, J. W. and Harris, C. W. "Reliability coefficients as a function of individual differences induced by a learning process assuming identical organisms." Journal of Mathematical Psychology, 1973, 10, Pp. 387-420.

Cronbach, L. J. Essentials of psychological testing. Third edition. New York: Harper and Row, 1970.

Cronbach, L. J. and Furby, L. "How should we measure 'change'--or should we?" Psychological Bulletin, 1970, 74 (1), Pp. 68-80.

Elashoff, J. D. "Analysis of covariance: a delicate instrument." American Educational Research Journal, 1969, 6, Pp. 383-401.

Evans, S. H. and Anastasio, E. J. "Misuse of analysis of covariance when treatment effect and covariance are confounded." Psychological Bulletin, 1968, 69, Pp. 225-234.

Glass, G. V., Peckham, P. D. and Sanders, J. R. "Consequences of failure to meet assumptions underlying the analysis of variance and covariance." Review of Educational Research, 1972, 42, Pp. 237-288.

Guilford, J. P. Fundamental statistics in psychology and education. Fourth edition. New York: McGraw Hill, 1956.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Heise, D. R. "Separating reliability and stability in test-retest correlation." American Sociological Review, 1969, 34, Pp. 93-101.

Jensen, A. R. "Estimation of the limits of heritability of traits by comparison of monozygotic and dizygotic twins." Proceedings of the National Academy of Science, April 26, 1967.

Kerlinger, F. N. Foundations of behavioral research. New York: Holt, Rinehart and Winston, 1964.

Lord, F. M. "Further problems in the measurement of growth." Educational and Psychological Measurement, 1958, 18, Pp. 437-454.

Lord, F. M. "A paradox in the interpretation of group comparisons." Psychological Bulletin, 1967, 68, Pp. 304-305.

Lord, F. M. "Statistical adjustments when comparing pre-existing groups." Psychological Bulletin, 1969, 72, Pp. 336-337.

Lord, F. M. and Novick, M. R. Statistical theories of mental test scores. New York: Addison-Wesley, 1968.

Marley, A. A. J. "Abstract one-parameter families of commutative learning operators." Journal of Mathematical Psychology, 1967, 4, Pp. 414-429.

Porter, A. C. "Comments on some current strategies to evaluate the effectiveness of compensatory education programs." Paper presented at the meeting of the American Psychological Association, 1969.

Rossi, P. H. and Williams, W. (Ed.) Evaluating social programs: theory, practice, and politics. New York: Seminar Press, 1973.

Stanley, J. C. "Analysis of variance of gain scores when initial assignment is random." Journal of Educational Measurement, 1966, 3, Pp. 179-182.

Stanley, J. C. "General and special formulas for reliability of differences." Journal of Educational Measurement, 1967, 4, Pp. 249-252.

Stanley, J. C. "Reliability." Chapter 13 in Thorndike, R. L. (Ed.), Educational Measurement, Second edition. Washington, D.C.: American Council on Education, 1971.

Thorndike, R. L. "Regression fallacies in the matched groups experiment." Psychometrika, 1942, 7, Pp. 85-102.