TEST CONSTRUCTION TECHNIQUES FOR BUILDING MORE SENSITIVE INDICATORS OF BETWEEN-GROUP DIFFERENCES

M. David Miller

CSE Report No. 174
1981

Center for the Study of Evaluation
Graduate School of Education, UCLA
Los Angeles, California 90024

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

While tests are used to assess the achievement differences between individuals, as well as ranking the achievement differences among aggregates of individuals, such as classrooms, schools or programs, the psychometric model used in the construction of norm-referenced tests has focused primarily on the former. The use of the individual as the unit of analysis in test construction combined with the largely negative results of school effects studies and large scale evaluations about the relationship of school inputs to pupil outcomes (Averch et al., 1972; Coleman et al., 1966; Jencks et al., 1972; Stebbins et al., 1977) has caused many educational researchers to reexamine the statistical techniques and models used to arrive at these conclusions.

A concern over the possible mismatch between the methods used to construct norm-referenced tests and the kinds of issues being addressed has led to questions about the program relevance and instructional sensitivity of norm-referenced measurement (Airasian & Madaus, 1976; Berliner, 1978; Carver, 1974; Hanson & Schutz, 1978; Leinhardt & Seewald, 1981; Madaus et al., 1979, 1980; Porter et al., 1978). This concern over the sensitivity of tests to instructional and program effects is evident in recent investigations of the overlap between test content and instructional content. These studies indicate that test performance is higher when there is substantial overlap between test content and instructional content (Armbruster et al., 1977; Jenkins & Pany, 1976; Leinhardt & Seewald, 1981; Madaus et al., 1979; Walker & Schaffarzik, 1974). This evidence in conjunction with the finding that there is wide variation in content coverage in the major standardized achievement tests (Porter et al., 1978) raises the question of whether schools are skilled at or successful at selecting the test that best fits their curriculum or whether this is even possible. Moreover, as long as teachers have the freedom to choose which topics to cover and emphasize within a subject area, tests may not be useful or relevant for measuring between-class differences.

Another concern about norm-referenced measurement has centered around the empirical methods used to construct tests. Some critics have argued that tests designed to differentiate among individuals can maximize the within-school differences relative to the between-school or between-program differences (Carver, 1974; Lewy, 1973). Theoretically, of course, there is no reason to assume that a test designed to measure individual differences cannot also measure school or program differences. However, the bulk of the evidence from school effectiveness studies suggests either that school or program differences are small or do not exist after controlling for home background and entering ability or that the between-group differences are not being measured properly (Madaus et al., 1980).

One approach to improving the sensitivity of measures of group differences might be to consider the inherent multilevel character of the educational system. That is students are nested within classrooms; classrooms are nested within schools, etc. Analyses can be conducted both between and within each of the levels of the educational system and analyses within and between the different levels can have different substantive meanings (Burstein, 1978, 1980; Burstein, Fischer,

& Miller, 1980; Cronbach, 1976). Thus, if analyses are not conducted from a multilevel perspective, one can fail to clearly identify important effects occurring at different levels. Because of a concern for the analyses of data at multiple levels, many major evaluations, such as Project Follow Through (Haney, 1974) and the National Day Care Study (Singer & Goodrich, 1979), have devoted considerable time and expense to the selection of the unit of analysis. Since education does affect student outcomes between and within all levels of the educational system, it has been argued that evaluations of educational data should look at more than one level of analysis for a more complete understanding of the determinants of student achievement. In fact, Cronbach (1976, p. 1) argued that the "majority of studies of educational effects -- whether classroom experiments, or evaluations of programs or surveys -- have collected and analyzed data in ways that conceal more than they reveal. The established methods have generated false conclusions in many studies."

While there has been a rapid rise in the concern for multilevel issues in large scale evaluations and school effects studies (see Burstein, 1980 for a review), most researchers have ignored the issue of multilevel data analysis in the construction of tests and the analysis of item data, with a few notable exceptions. In his monograph on multilevel issues, Cronbach (1976, p. 9.19-9.20) discussed the possible utility of multilevel item analysis:

> Once the question of units is raised, all empirical test construction and item-analysis procedures need to be reconsidered. Is it better to retain items that correlate across classes? Or items

> that correlate within classes? A correlation based on deviation scores within classes indicates whether students who comprehend one point better than most students also comprehended the second point better than most -- instruction being held constant. A correlation between classes indicates whether a class that learned one thing learned another, but this depends first and foremost on what teachers assigned and emphasized. It is the items teachers give different weight to that have the greatest variance across classes. This (differential emphasis) leads us to regard the between-group and within-group correlations of items as conveying different information, and makes the overall correlation for classes pooled an uninterpretable blend.

As Cronbach (1976) suggests, it may be useful to reexamine the empirical techniques used in item analysis and test construction in a multilevel contest. Hence, instead of using indices of item discrimination between subjects in test construction, indices of item discrimination between groups may prove more useful in building scales more sensitive to differences between groups. One test construction technique for building tests more sensitive to between-group differneces was suggested by Lewy (1973). Since the purpose of the test is to discriminate between groups, Lewy suggested an index of how well items discriminate between groups as a criterion for inclusion in the test - the intraclass correlation. The intraclass correlation is equal to the proportion of variation in an item that is attributable to group differences. Thus, the intraclass correlation coefficient equals one when all scores within each group are identical and the only variance is due to differences between groups. Conversely, the intraclass correlation coefficient equals zero when all the group means are equal and the only variation is due to differences within a group. Lewy proposed the intraclass coefficient to be used to identify subsets of items that maximize the variance between-groups on the total test relative to the total test score variance.

While the intraclass correlation coefficient may be a useful index of how well an item differentiates between groups, using it as the sole criterion for item selection may be overly simplistic. As Airasian and Madaus (1976) point out, items may differentiate between groups in different directions, so that they fail to discriminate between groups when summed into a single composite. For example, given two groups of equal size, if everyone in one group answered one item correctly and the other item incorrectly, while the reverse was true for the second group, then the two items would each have an intraclass correlation of one, but the sum of the items would not discriminate between groups at all. Because of this phenomenon, Airasian and Madaus suggested using the between-group intercorrelations of the items along with the intraclass correlations. It could even be argued that the intercorrelations are a more important piece of information since the variance of an n-item scale is equal to n item variances and n(n-1) item covariances. So the between-group item intercorrelations could be used to develop a scale which maximized the variance between groups.

Using the item between-group intercorrelations will create a scale which is internally consistent for discriminating groups. This procedure can rapidly become unwieldly, however, since there are n(n-1)/2 intercorrelations between n items. Because of this, a procedure that has been used to build internally consistent scales for measuring individual differences might also be applied to build an internally consistent scale for measuring differences between groups. That is, rather than the point-biserial correlation between the items and the total scale, the correlation between the wieghted item group means and the group means on the total scale could be used. Thus, the information needed for any decisions is reduced from n(n-1)/2 to n.

One final approach to item selection would be to use some criterion external to the test for item selection. For example, the Beginning Teacher Evaluation Study (BTES) had some success in developing scales sensitive to instructional differences between individuals (BTES: Filby & Dishaw, 1975, 1976). However, in the BTES study, all instructional variables were measured at the student level (e.g., allocated time). Because this is not always possible due to practical considerations (e.g., the time and expense that would be needed in a larger study), as well as the fact that many instructional variables cannot be measured at the student level (e.g., number of aides or money invested), the criteria used in item selection might be group-level measures (e.g., instructional materials) or even aggregate measures of individual-level variables (e.g., opportunity to learn). Even when the individual-level measures of the instructional variables (e.g., instructional time) are available for the item tryout, the relationship of the items to the aggregate measure might be used for item selection, if the unit of analysis is the aggregate in the final study.

Data Analysis

Sample. The Beginning Teacher Evaluation Study (BTES: Fischer et al., 1978) was sponsored by the California Commission for Teacher Preparation and Licensing with funds from the National Institute of Education. The study was conducted to examine the relationship of reading and mathematics achievement to instructional variables in grades 2 and 5. Fractions was a subject area in which a great deal of time and effort are expended in many fifth grade classrooms. Tests were administered to six students in each of 25 second and 25 fifth grade classes on four occasions -- (A) October, 1976; (B) December, 1976; (C) May, 1977; and (D) September, 1977. Since there was very little fraction instruction until after December, the October testing did not include the fractions subtest. In addition to the achievement tests, measures of allocated time, engagement rates and success rates were obtained. Also, teacher behavior measures were collected. To reduce the variability due to initial ability and home background, students were not selected who scored extremely low or extremely high on a selection test at the beginning of the year. Selected students were roughly between the 30 and 70 percentile of the overall distribution from all classes.

The fraction subtest data consisted of fifteen items administered on three occasions. The skills tested included fraction addition, fraction subtraction, reducing fractions and finding the missing numerator or denominator in a fractional equation. Data was obtained from 127 students on occasion B (December, 1976), 123 students on occasion C (May, 1977), and 89 students on occasion D (September, 1977). The students were drawn from 21 classrooms.

In addition, the pilot data will be used for the test construction. Because of an interest by the BTES in instructional variables, special effort was made to develop instructionally sensitive measures (BTES: Filby & Dishaw, 1975, 1976). Two criteria were used to enhance the likelihood that the tests would be instructionally sensitive. First, item content was checked to be sure that instructional content and test content overlapped. Next, items were checked to see if gains in achievement were related to gains in instruction (Carver, 1974). This second criterion involved two assumptions. First, students would perform better after instruction than before instruction. Second, students who receive more instruction would achieve higher than students who received less instruction. Consequently, the pilot study, conducted in April, 1975, included both test item data and a measure of allocated time. The sample included 72 subjects drawn from 6 classrooms.

Data Analysis. Three of the item selection techniques outlined above will be used to form subscales of the fraction test. Items will be selected on the basis of their characteristics in the spring testing of the pilot study and the corresponding scale will be examined in the spring testing of the final BTES study. The three criteria used in item selection are:

(1) the ability of the item to discriminate between groups by itself (i.e., intraclass correlation);

(2) how the item discriminates in relationship to the total scale (e.g., correlation of class means on items and class means on total scale); and,

(3) whether the item discriminates between classes that vary in instruction (e.g., correlation of class means on items and class means on allocated time in fraction instruction).

The primary criterion used to judge the utility of these test construction methods will be the intraclass correlation of the formed scale. However, when the correlation of the mean allocated time and the item means by classroom is used for item selection, the resulting scale's relationship to instructional variables in the final study will also be examined.

Results and Discussion

Three properties of the items were used to form scales. The first criterion was the proportion of variance in the item attributable to the differences between classes - the intraclass correlation. The second criterion was the relationship of the item to the total scale - the correlation of the class means on the item with the class means on the total scale. The third criterion was the relationship of the item to another variable - the correlation of the class means on the item with the class means on time allocated to fractions. The descriptive statistics used in the item selection are contained in Tables 1, 2, and 3.

The intraclass correlation for the fifteen item scale in the final study was .47. Forming scales from the item intraclass correlation did not increase the ratio of between class variance to total variance between subjects. Selecting the ten items with an intraclass correlation greater than or equal to .10 or the four items with intraclass correlations greater than or equal to .15 led to an intraclass correlation on the scale of .46 and .44, respectively. Similarly, selecting items so that the between-class item-total scale correlation was greater than or equal to .75 and .80 led to scales with an intraclass correlation of .45 (9 items) and .42 (5 items), respectively. Finally, selecting the four items with a positive between-class correlation of allocated time and the item ($\rho \geq .05$) led to a scale with an intraclass correlation of .42. Hence, selecting items on the basis of their statistical properties does not seem to increase the proportion of variance in the scale that is due to group differences.

However, selecting items on the basis of their relationship to another variable did increase the sensitivity of the scale to the variable of interest. In Table 4, the fifteen item scale and the four item scale formed by the between-class correlation of mean allocated time and the item means are predicted from the same set of variables - the pretest, allocated time, engagement rate, hard time, and easy time. By examining the standardized regression coefficients, it can be seen that the greatest differences in the prediction of the two scales is in their sensitivity to two variables which are similar to the criterion used in item selection -- allocated time and engagement rate. By constructing the scale on the basis of the between-class relationship of the items to instruction, the sensitivity of the scale to between-class differences in instruction is increased and the sensitivity of the scale to the same two variables within-class is decreased. Thus, if the object is to determine the relationship of achievement to differences between classes in instruction, it may be useful to select the items for the achievement test on the basis of their relationship to the variable of interest.

References

Airasian, P.W., and Madaus, G.F. A study of the sensitivity of school program effectiveness measures. Report submitted to the Carnegie Corporation, New York. Chestnut Hill, Mass.: Boston College, School of Education, 1976.

Armbruster, B.B., Steven, R.O., and Rosenshine, B. Analyzing content coverage and emphasis: A study of three curricula and two sets. Technical Report No. 26, Center for the Study of Reading, University Illinois, Urbana-Champaign, 1977.

Averch, H. Carroll, S.J., Donaldson, T., Kiesling, H.J., and Pincus, J. How effective is schooling? A critical review and synthesis of research findings (R-956-PCSF/RC). Santa Monica, CA: The Rand Corporation, 1972.

Berliner, D.C. Studying instruction in the elementary school classroom: Clinical educational psychology and clinical economics. Paper commissioned by the Education and Finance and Productivity Center, Department of Education, University of Chicago, 1978.

Burstein, L. Assessing differences between group and individual-level regression coefficients. Sociological Methods and Research, 1978, 7(1), 5-28.

Burstein, L. Analysis of multilevel data in educational research and evaluation. In D. Berliner (Ed.), Review of research in education, Vol. 8. Washington, D.C.: American Educational Research Association, 1980, 158-233.

Burstein, L., Fisher, K., and Miller, M.D. The multilevel effects of background on science achievement at different levels of analysis: A cross-national comparison. Sociology of Education, 1980, 58(4), 215-255.

Carver, R.P. Two dimensions of tests: Psychometric and edumetric. American Psychologist, 1974, 29, 512-518.

Coleman, J.S., Campbell, E.Q. Hobson, C.J. McPartland, J., Mood, S., Weinfeld, F.D., and York, R.I. Equality of educational opportunity (2 Vols.). Office of Education and Welfare. Washington, D.C.: U.S. Government Printing Office, 1966.

Cronbach, L.J. (with the assistance of J.E. Deken & N. Webb). Research on classroom and schools: Formulation of questions, design, and analysis. Occasional Paper, Stanford Evaluation Consortium, Stanford, California, July, 1976.

Filby, N.N., and Dishaw, M. Development and refinement of reading and mathematics tests for grades 2 and 5. Technical Report III-1, Far West Laboratory for Educational Research and Development Beginning Teacher Evaluation Study, August 1975.

Fisher, C.W., Filby, N.N., Marliave, R.S., Cahen, L.S., Dishaw, M.M., Moore, J.W., and Berliner, D.C. Teaching behaviors, academic learning time and student achievement: Final Report of Phase III-B, Beginning Teacher Evaluation Study (Technical Report V-1). San Francisco, CA: Far West Laboratory for Educational Research and Development, June 1978.

Haney, W. Units of analysis issues in the evaluation of Project Follow Through. Unpublished report, Cambridge, Mass.: Huron Institute, 1974.

Hanson, R.A., and Schutz, R.E. A new understanding of schooling effects derived from programmatic research and development. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, April 1978.

Jencks, C.S., Smith, M., Acland, H., Bane, M.J., Cohen, D., Gintis, H., Heyns, B., and Michelson, S. Inequality: A reassessment of the effect of family and schooling in America. New York: Basic Books, 1972.

Jenkins, J.R., and Pany, D. Curriculum biases in reading achievement tests. Technical Report No. 16. Urbana, IL: University of Illinois, Center for the Study of Reading, November 1976.

Leinhardt, G., and Seewald, A. Overlap: What's tested, what's taught? Journal of Educational Measurement, in press.

Lewy, A. Discrimination among individuals vs. discrimination among groups. Journal of Educational Measurement, 1973, 10, 19-24.

Madaus, G.F., Airasian, P.W., and Kellaghan, T. School effectiveness: A reassessment of the evidence. New York, NY: McGraw-Hill Book Company, 1980.

Madaus, G.F., Kellaghan, T., Rakow, E.A., and King, D.J. The sensitivity of measures of school effectiveness. Harvard Educational Review, 1979, 49(2), 207-230.

Porter, A.C., Schmidt, W.H., Floden, R.E., and Freeman, D.J. Practical significance in program evaluation. American Educational Research Journal, 1978, 15(4), 529-539.

Singer, J.D., and Goodrich, R.L. Aggregation and the unit of analysis in the National Day Care Study. Paper presented at the Annual Meeting of the American Research Association, San Francisco, CA, April 1979.

Stebbins, L.B., St. Pierre, R.G., Proper, E.C., Andersen, R.B., and Cerva, T.R. Education as experimentation: A planned variation model. Cambridge, Mass.: Abt Associates, 1977.

Walker, D.F., and Schaffarzick, J. Comparing curricula. Review of Educational Research, 1974, 44(1), 83-111.

Table 1. BTES item intraclass correlations (η^2) on on the spring pilot testing.

Item	η^2	Item	η^2	Item	η^2
1	.14	6	.03	11	.11
2	.11	7	.16	12	.11
3	.08	8	.18	13	.24
4	.11	9	.06	14	.05
5	.04	10	.18	15	.10

Table 2. BTES between-class item-total (ρ) correlations on the spring pilot testing.

Item	ρ	Item	ρ	Item	ρ
1	.91	6	.94	11	.79
2	.88	7	.54	12	.45
3	.97	8	.83	13	.56
4	.77	9	.77	14	.46
5	.66	10	.75	15	.45

Table 3. BTES between-class correlation (ρ) of the item and time allocated to fractions from the spring pilot study.

Item	ρ	Item	ρ	Item	ρ
1	-.57	6	-.35	11	.01
2	-.53	7	.16	12	-.05
3	-.70	8	-.20	13	-.42
4	.51	9	.19	14	-.51
5	.45	10	-.64	15	-.49

Table 4. Prediction of the spring achievement test and a subscale from the pretest and instructional variables from the BTES final study.[a]

	Total Scale		Items 4, 5, 7, & 9	
Within Class	Unstandardized	Standardized	Unstandardized	Standardized
Pretest	.39 (11.78)	.34	.16 (17.01)	.43
Allocated Time	.01 (1.52)	.16	-.00 (.06)	-.03
Engagement Rate	2.50 (1.94)	.16	-.11 (.03)	-.02
Hard Time	-11.76 (1.01)	-.16	-3.51 (2.89)	-.14
Easy Time	2.44 (3.88)	.11	.59 (.50)	.08
Between Class				
Pretest	.24 (.45)	.14	.02 (1.52)	.03
Allocated Time	.02 (.67)	.24	.01 (1.52)	.38
Engagement Rate	1.07 (.05)	.05	1.87 (1.15)	.27
Hard Time	-1.91 (.06)	-.06	.68 (.06)	.07
Easy Time	-8.74 (.10)	-.06	-6.23 (.43)	-.12
R^2	.52		.47	

[a] F-tests in parenthesis.