STUDENT ENTRY SKILLS AND THE EVALUATION OF INSTRUCTIONAL PROGRAMS: A CASE STUDY

Rodney W. Skager

CSE Report No. 53
June 1969

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

STUDENT ENTRY SKILLS AND THE EVALUATION OF INSTRUCTIONAL PROGRAMS: A CASE STUDY

It is axiomatic among curriculum experts that teachers often fail to acquaint themselves with the patterns of skills students bring initially to their classrooms. When new instructional programs are being developed for later use by large numbers of teachers, such failure to monitor entry skills can result in a grossly inadequate match between the learning needs of students and the content of instruction. The present research goes further, however, and suggests, that under certain circumstances there may even be a tendency for teachers to emphasize skills already mastered by their students. This paper presents empirical evidence for such an assertion and attempts to deduce some of the reasons why teachers under conditions similar to those encountered in this research might direct instruction at the improvement of skills already attained by a majority of their students.

Students and Instructional Programs

Findings reported in this research are based on data collected as part of the evaluation of a curriculum development program in seventh grade mathematics. For the analyses reported here data were available on 488 students in three junior high schools. Of these, 285 were assigned to experimental classes taking the curriculum under development, and 203 were assigned to comparison classes providing the regular mathematics curriculum for each school. Within each school experimental and comparison groups were not in every case equivalent at the beginning of instruction, but this fact has no bearing on the issues discussed here. The three schools were located in a metropolitan context, and in two cases were of an "inner" city type. Students were varied in racial and ethnic characteristics. As identified below, approximately 95% of the students in School 1 were Mexican-Americans, with an approximately equal percentage of the students in School 2 being Negro. School 3 was of a more mixed ethnic character, with somewhat over 60% of the students Caucasian, about 30% Mexican-American, and a small percentage of Negro and Oriental students. The two sexes were approximately equally represented in all groups.

The majority of the students in the three schools had taken the California Test of Mental Maturity (1957 Short Form) at the end of the fifth grade. Mean total I. Q. scores at each school for experimental and comparison groups were respectively: School 1, 94 and 90; School 2, 90 for both groups; and School 3, 95 and 100. All students assigned by the schools to either experimental or comparison classes were at least one year behind in mathematics achievement for their own school. In general, then, students participating in this research can be described as members of urban minority groups who have shown unsatisfactory achievement in mathematics in comparison with their peers. Their academic performance probably cannot be accounted

The teachers were instructed to use the following 5-point rating scale:

1. Definitely would not facilitate ability to answer
2. Probably would not facilitate
3. Uncertain
4. Probably would facilitate
5. Definitely would facilitate ability to answer

Mean ratings for teachers in each combination of test form, school, and instructional program were calculated for each item.

Purpose

Initially, the relevancy ratings were collected to provide a check on the fairness or appropriateness of the items included in the Diagnostic Test. It was hoped that all or nearly all items would be judged to be closely related to the content of instruction, especially in the experimental group. It was also of interest to determine whether the comparison teachers would see the items as less relevant than did the experimental teachers, as might be expected in view of the fact that their instructional goals were not specific criteria used in selecting the items.

In examining the data, however, the author chanced to see an initial item difficulty index for one of the subgroups in juxtaposition with the mean rating of the item by the teachers of those particular students. All of the students in the group had passed the item at pretest, yet all of the teachers of those students had rated the item 5, implying a definite relevancy to instructional content! This startling observation led to investigation of the overall relationship between initial item difficulty and teacher ratings of item relevancy. For this purpose, correlations between proportion of students passing each item at pretest and mean rating of item relevancy were calculated for each subgroup of students. Ideally, such correlations should be negative, indicating that teachers place greater emphasis on those skills in which students are initially weak. Correlations of approximately zero would suggest the lack of any systematic relationship between entry skills and instructional content, seemingly an undesirable situation. Positive correlations, of course, would be even less desirable, since such findings would suggest that instructional content is oriented to student strengths rather than weaknesses.

Results

As indicated above, the data collected in this research were analyzed for the purpose of determining the nature of the relationship between the entry skills of students and the instructional objectives of their teachers. Before presenting evidence relating to this primary issue, two preliminary questions need to be dealt with by way of anticipating possible alternative interpretations of the results.

(1) Did the teachers in both groups judge items on the Diagnostic Test to be in general relevant to their instructional objectives and were there differences in the ratings of experimental and comparison teachers? Mean ratings averaged over the 52 items in

each form of the pretest are reported in Table 1. These ratings were also identified according to school and experimental vs. comparison teachers, and they show that as a whole the test was judged relevant to instructional goals as perceived by the teachers themselves. All but one of the means are above 3.0, the point of uncertainty. Perusal of the data on individual items for each of the twelve subgroups shown below did show variability for all groups in ratings across items, but also revealed many items with mean ratings at or close to 5.0 for a given group.

It was anticipated that the experimental teachers would judge the test to be more relevant, since their planned instructional goals were the major consideration in the selection of test items. This expectation was not confirmed. The mean ratings in Table 1 do not reveal a pattern of differences between ratings by experimental and comparison teachers. In most cases the means are very close for the two groups within each school, and the highest mean in the table was generated by comparison teachers at School 1. (2) To what extent did experimental and comparison teachers agree on the relevancy of individual items of the test? While overall ratings of the relevancy of items were remarkably similar for the experimental and comparison teachers, it does not necessarily follow that teachers in the two groups saw the same items as relevant. Indeed, if all teachers gave similar relevancy ratings on each item, any differences between the two groups in the correlations of initial item difficulty with relevancy ratings could only be explained in terms of systematic differences between experimental and comparison students in the skills available at entry. Since there is no reason to believe that the process of assigning students to groups would result in systematic differences in patterns of entry skills, this explanation of the results would lead nowhere.

To answer the above question, mean ratings by experimental and comparison teachers on each item were correlated over the 52 items

Table 1

Mean Ratings of Relevancy of Diagnostic Test Items*

	Form A		Form B	
	Experimental	Comparison	Experimental	Comparison
School 1	3.71 (5)	4.28 (3)	3.81 (5)	4.23 (3)
School 2	4.0 (4)	3.94 (8)	3.98 (4)	3.82 (8)
School 3	3.61 (4)	3.59 (2)	3.36 (4)	2.95 (2)

* Numerals in parentheses refer to number of teachers contributing to each mean rating. The rating for item relevancy ranged from "1" (low) to "5" (high).

within school and by test form. These correlations are reported in Table 2. Inspection of these correlations reveals that only in the case of School 2 is there a relatively high relationship between the relevancy ratings of experimental and comparison teachers. In the case of the other two schools, the correlations, while positive, are quite weak. A possible interpretation of this finding may lie in the report made by members of the evaluation staff assigned to the schools as periodic observers that only at School 2 were either formal or informal discussions between experimental and comparison teachers about instructional objectives known to have occurred. Although the Diagnostic Test appears to be based on reasonably appropriate overall content for both experimental and comparison classes, it appears that different subgroups of items were seen as relevant by experimental and comparison teachers in two of the schools, with moderate positive relationships at the third school. With the above in mind the primary question posed in this paper can be addressed. (3) What relationship pertained between the entry skills of students as reflected in initial item difficulty and ratings by teachers of the instructional relevancy of those items? Correlations between initial item difficulty across the 52 items on each form are reported in Table 3 for each of the twelve subgroups. Two trends are immediately apparent in this table. First and most important, all of the correlations are positive, confounding the seemingly reasonable expectation that the signs of the coefficients would be negative. Table 3 reveals very clearly that the larger the proportion of students able to answer each item correctly at the beginning of the year, the more likely were teachers to rate that item as highly relevant to their instruction. In short, by their own reports, the teachers appeared to have selected instructional objectives that to a considerable extent reflected skills already available to their students.

Table 2

Correlations Between Mean Relevancy Ratings on Individual Test Items for Experimental and Comparison Teachers

	School 1	School 2	School 3
Form A	.35	.62	.20
Form B	.12	.64	.18

Table 3

Correlations Between Proportion Answering Each Item of Diagnostic Test Correctly at Pretest and Teacher Ratings of Item Relevancy

	Form A		Form B	
	Experimental	Comparison	Experimental	Comparison
School 1	.25	.14	.62	.30
School 2	.44	.27	.58	.32
School 3	.53	.10	.62	.25

The second trend is also surprising. Without exception correlations are higher for experimental groups than for corresponding comparison groups for each combination of school and test form. For Form A the average $\bar{r}$ for experimental groups across schools is .42 as compared to .17 for comparison classes. For Form B the average experimental group correlation is .61 as against .29 for comparison students. There thus appears to have been a greater tendency among experimental teachers to gear instruction to skills already achieved by students at entry into the program.

It may also be noted, incidentally, that the correlations in Table 3 are invariably higher for Form B of the test. This trend can probably be ignored, since the author neglected to control for order effects when the ratings were collected, with the result that the items on Form A were always rated first. The correlations for Form B are perhaps more accurate estimates in the sense that the judges were more practiced.

Discussion

How are these results, so inconsistent with what seems to be a reasonable expectation, to be explained, and what are their implications for the development and evaluation of instructional programs? We are, of course, dealing here with correlational research designed to identify relationships existing in the data rather than to explain, as would be the case under experimental conditions, the origin of relationships. For this reason, and because of the possible importance of these findings with regard to educational practice, several alternative explanations need to be considered.

A first explanation deserving of consideration holds that the results reported above on the relationship between entry skills and relevancy ratings are spurious in the sense that the teachers could actually have emphasized different content in the classroom than was indicated by their ratings. That is, perhaps the ratings did not reflect what the teachers actually did, but rather the opposite or at least something quite different. Admittedly, the motivation for such behavior is difficult to construe, but reasoning along the following lines does not appear unduly contrived. We can safely assume that teachers are sensitive about evaluations others make of their performance as reflected in the achievement of their students. Moreover, teachers have sufficient opportunity during the year to become aware of the patterns of subject matter skills available to their students. Given this combination of desire to "look good" and knowledge of what students can and cannot do, it would be easy to claim credit via the relevancy ratings for teaching students what they already knew in the first place.

While such an uncharitable interpretation of the results cannot be completely discounted, it does seem improbable for at least two reasons. First, there is no positive evidence for the assertion that teachers were either consciously or unconsciously distorting the actual situation in their ratings. On the contrary, there is some evidence partly formal and partly informal that the ratings were honest reflections of instructional content. In another report derived from this same research Patalino (1968) found frequent instances in which greater than average gains from pretest to posttest were accompanied by higher than average relevancy ratings for subgroups of items examined separately, suggesting that more emphasis was placed on skills rated highly relevant. The subgroups of items were not formed on the basis of an analysis of the teachers' ratings (as had originally been intended) but rather because of judged similarity of content. There are also informal instances of reports by observers of students commenting to the teachers that at least some of the material was familiar.

A second interpretation of the results assumes that the ratings do reflect accurately the content emphasized by the teachers. This assumption is at least consistent with the tentative evidence just cited. Again given that they are motivated to be judged effective in their work, this interpretation asserts that teachers find it tempting to teach available skills, knowing consciously or unconsciously that their students will then appear to be performing well, especially when they are being observed by outside evaluators. This explanation would account for the differences between the correlations in Table 3 for experimental and comparison groups in the sense that the experimental teachers were undoubtedly under greater internal pressure to succeed, since they were "master" teachers participating in an experimental program with high visibility. Not only behavioral scientists but also a variety of educators were observing their classroom and materials. Except for the achievement testing, comparison teachers were in a very much more typical situation with regard to visibility.

The above explanation is plausible, but there is an additional interesting possibility. The instruction of urban,

minority children who are not achieving as well as their own peers is likely to be very hard work for teachers. Moreover, experimental teachers participated in sensitivity training sessions in which it was stressed that such children are likely to associate academic aspects of school with a sense of personal failure and inadequacy. In effect, experimental teachers were thus being urged not to give the children in the program further experiences with failure. These two conditions would also account for the fact that both experimental and comparison teachers apparently directed instruction at skills already available (it was an easier alternative than trying to teach new content), as well as the fact that experimental teachers did so to a greater degree in order to avoid confronting their students with further failure experiences.

As plausible as the two explanations may be for the conclusion that the relevancy ratings reflected instructional content accurately (and both could be operating at the same time), one could still argue that the results do not make sense because learning would not go on at all in the schools if instructional objectives were confined to what students already knew. In reply it can be noted that while the above correlations are not perfect relationships and do not completely exclude the possibility of some new material being introduced into the curricula studied, as it undoubtedly was, this report does not deal with students from the affluent middle classes, but with urban minority students who are already far behind in achievement and who, if Coleman's (1966) findings apply, will fall further behind with time. The present results are quite consistent with this well-documented phenomenon. Thus, it seems reasonable to conclude that the frustrations encountered in teaching educationally handicapped students as well as the need perceived by teachers to provide such students with experiences of success and also the teachers' own needs to perform well, especially under conditions of close observation, may well lead teachers to make the task of instruction easier by emphasizing those areas of content in which present capabilities of students are relatively more developed.

Implications

Of the two major implications of these findings, the first relates to instructional practice and the second to the methodology of evaluation. With regard to the former, it is readily apparent that teachers do need to be informed about the entry skills of their students as related to the objectives of a course of instruction, because without information there is undue latitude for the operation of other irrelevant factors in decisions about curriculum. Such information on entry skills will be most useful if referred to specific, unequivocal objectives such as those described by Popham (1969). The importance of obtaining information on entry skills has, of course, been stressed by others, including Glaser (1967), in the development of individualized instructional curricula.

Secondly, it is clear in the present case that data on entry skills could have been most useful to the teachers developing the program had they been made available early in the research. This failure to meet the needs of program developers is illustrative of an all too common phenomenon in evaluation. There is a widespread

tendency for researchers engaged in the evaluation of instruction to concentrate on collecting data relevant to program adoption at the expense of data relevant to program development. That is, behavioral scientists typically approach evaluation of educational practices with the analogue of the experiment firmly in mind. This leads to undue concern with answering the question, "Is the new program better than the old?", and results in neglect of the more important task of helping program developers make certain the answer will turn out to be in the affirmative. Unlike the experimenter in a controlled laboratory situation, it is highly appropriate for the educational researcher in the role of evaluator to produce data that will lead to modifications in "treatment" variables while the research is going on. As illustrated in the present case, if the evaluator does not seek out systematic information relevant to program development, it is not likely that others will. This need has already been pointed out by Cronbach (1963), Stufflebeam (1968), and by others. Certainly, the present research provides strong support for the assertion that all who are involved in the development and evaluation of programs of instruction should monitor the entry skills of the target population.

REFERENCES

Coleman, J. S. Equality of educational opportunity. United States Office of Education, 1966.

Cronbach, L. J. Course improvement through evaluation. Teachers College Record, 1963, 64, 672-683.

Glaser, R. Adapting the elementary school curriculum to individual performance. Proceedings of the 1967 Invitational Conference on Testing Problems, Educational Testing Service, 1967, 3-36.

Patalino, M. The rationale and use of content relevant achievement tests for the evaluation of instructional programs. Unpublished Master's Thesis, Graduate School of Education, University of California, Los Angeles, 1968.

Popham, W. J. The controlling effect of educational objectives on curriculum. Encyclopedia of Education, in press.

Stufflebeam, D. L. Evaluation as enlightenment for decision-making. Evaluation Center, Ohio State University, Columbus, Ohio, 1968.