DOCUMENT RESUME

ED 052 530 EA 003 599

AUTHOR Rapp, M. L.; And Others
TITLE Some Considerations in the Experimental Design and Evaluation of Educational Innovations.
REPORT NO P-4360
PUB DATE Apr 70
NOTE 13p.

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Cost Effectiveness, Criteria, *Decision Making, *Educational Innovation, *Evaluation, *Experiments, Input Output Analysis, Models, *Program Design, Program Development, Student Characteristics
IDENTIFIERS Program Selection Criteria

ABSTRACT

The parameters of an evaluation design determine the use to which the evaluation is to be put, the ultimate user of the results, and the capabilities of the school information system. Evaluation supports decisionmaking in program adoption and program improvement, as well as research for a better understanding of the educative process. An experimental design should, therefore, be structured to accommodate all data requirements for the evaluations. Planning for future implementation should be concurrent with the planning for innovative programs. (Author)

SOME CONSIDERATIONS IN THE EXPERIMENTAL DESIGN AND EVALUATION OF EDUCATIONAL INNOVATIONS

M. L. Rapp
J. G. Root
G. Sumner

April 1970

P-4360

SOME CONSIDERATIONS IN THE EXPERIMENTAL DESIGN AND EVALUATION OF EDUCATIONAL INNOVATIONS

M. L. Rapp*
J. G. Root
G. Sumner

The RAND Corporation, Santa Monica, California

I. INTRODUCTION

The evaluator's task is to relate inputs (student and school characteristics) to outputs (cognitive or affective changes). It is generally accepted today that a student's performance depends on characteristics related to his home life, such as his abilities and attitudes, and those of his parents, his racial identity and socio-economic status--and on characteristics related to his school environment--his teachers and other school personnel, the curriculum and facilities available to him.(1)

The use to which the evaluation is to be put, the ultimate user of the results, and the capabilities of the school information system, determine the parameters of the evaluation design. There are three basic functions served by evaluation: to support decisionmaking in program adoption, to support decisionmaking in program improvement, and to support research for a better understanding of the educative process.

Evaluation results that specify what achievement gain is being produced by what resource mix for which segment of the population, can be combined with a cost analysis of the program resources to provide the essential ingredients for a cost/effectiveness analysis.(2) This information can then be used to aid the decisionmaker in choosing among alternative programs.

---

*Any views expressed in this Paper are those of the authors. They should not be interpreted as reflecting the views of The RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The RAND Corporation as a courtesy to members of its staff.

Evaluation for program improvement is concerned with the details of each program and the results are used within the program on a short-term basis to improve the operation of the program in meeting its stated goals. In this role the evaluator becomes the focal point in a feedback loop, garnering information about the effectiveness of the program that can then be used to improve program design.(3)

In addition to these action-oriented uses for evaluation results, the researcher can use them in conjunction with other results to improve his understanding of the educational process. Then, over a longer time span, the research results can be fed back to decisionmakers to aid decisionmaking in initial program selection and continuing program improvement.

A precondition for evaluation is the specification of the program to be evaluated, the relevant effectiveness measures for the program, and the types of students on whom the program is to be tested. Because decisionmakers at various administrative levels are faced with different problems, their interests as to programs, measures, and students will differ. Since the results of the evaluation need to be pertinent to school decisionmaking at a number of levels, the evaluators are faced with a multiplicity of tasks.

The evaluator translates the above specifications into an evaluation model that may vary from the simple to the complex. The complexity of the model depends, of course, on the number and kinds of programs, students, and effectiveness measures that are to be considered. The evaluator's model describes the manner in which the inputs are assumed to effect the outputs (e.g., achievement).

The type of model adopted depends somewhat on how the evaluation results are to be used. For example, if they

are to be used for program improvement, then the model should reflect the details of the program so that the effect of each can be determined and program changes made where necessary. On the other hand, a detailed model may not be required if to support program adoption one simply wants to know which of two programs is more effective for a standardized mix of students. Unfortunately, the payoff from using a simple model that does not place much emphasis on understanding the educational process is correspondingly small. The use of simple models has dominated evaluation and research efforts for many years and has failed to produce much knowledge that could be used to improve the educational process.

The more emphasis that is placed on evaluation for understanding the educational process, the larger is the requirement for a more varied mix of inputs, for more detailed measurements, for a more complex analysis, and thus, for a better designed school information system. Also, if the evaluation results are to be used for program improvement, then the information system must be able to make faster responses to information requests.

Because the school information system will largely be shaped by the requirements for evaluation and resource analysis, it should be designed to meet these needs. Information systems designed in a vacuum, generally from readily available data, rarely, if ever, support the data requirements of the decisionmaker.

II. CONSIDERATIONS IN THE EXPERIMENTAL DESIGN

Experimental design is essentially a problem of organizing the observation of various alternatives and of specifying criteria and instruments of measurement. There are times when evaluators must face the problem of compromising tenets of "good" experimental design to accommodate the realities of implementation (political, economic, and social realities).

The experimental design should be so structured as to accommodate all the data requirements for the evaluations. Often, one bit of information will serve several evaluation purposes. For instance, reading-achievement gain may be the criterion for several programs. For any one program it is a measure of the effectiveness of a specific set of educational inputs in achieving an objective. It can also be used to compare the amount of gain to be expected from several different mixes of inputs. Finally, if it is also related to student and teacher characteristics, it serves to further an understanding of the variables that contribute to the achievement of sub-sets of the student population.

Throughout the life of the research effort, objectivity requires adherence to the experimental design. However, decision rules should be incorporated that allow midstream course changes; rigid implementation of impotent programs (as well as frequent program changes) may do injustice to the cost-effectiveness of overall research, not to mention the students taking part in the program.

Experimental Control. To the extent that the variables to be measured are affected by inputs other than the program alternatives, the design should provide control on those inputs. Control accomplishes two purposes: (1) it increases the representativeness of the experiment with respect to those inputs, in turn enhancing generalizability of experimental results to the target population; (2) it allows the analyst to test for interaction between program inputs and non-program inputs (e.g., a program input may be effective for children of certain backgrounds, but not for others). Where there is no interaction, the problem of generalizing from innovative programs to all schools will require far less heroism because representation over the non-interactive controls is that much less important. In addition, the presence or absence of interaction is useful information about the educative process.

To control on a variable is to apply each competing program alternative under each of several levels of that variable, e.g., students from each of several socio-economic levels, teachers with different training, or districts with different administrative policies. In cases where an experimental program provides a complementary function to another experimental program one may be treated as a control on the other and vice versa.

The fixed-level budget for any experimental program requires that a selection be made with regard to the number of possible combinations of inputs and levels of inputs. It is possible that economic expediency will preclude controlling an experiment by some of the important non-program variables. In this case, one may at least want to set up control groups for the program alternatives. The control group is a handy expedient when there are insufficient resources to formally incorporate non-program effects into the design, or when there is uncertainty as to which non-program effects are important. Setting up a control group involves designating a class or school that embodies all characteristics (student, teacher, curriculum, etc.) of the experimental class or school, except that the purposes of the experimental program are met therein by "conventional" means. A control group provides an experiment with only very limited generalizability, answering little more than whether the particular students in the experiment might have done better or worse without the special program.*

To some extent, the benefits of control groups, or of controlling on individual variables, are tentative. Educational experiments lack the relative homogeneity

*Even if controlling by individual variable is possible, control groups may be desirable in order to "standardize" achievement-measuring methods used in experimental programs (especially if those methods bear little commonality with conventional tests).

of experimental conditions encountered in the biological and physical sciences; it is in fact difficult to even approximate the ideal of setting up independent experimental trials where some inputs are held constant and other inputs are varied in a known manner. Careful planning is necessary to minimize some of the more obvious sources of this experimental "looseness."

One such source, for example, arises from the fragmented nature of many innovative programs. Children often participate in a "standard" district program for part of a day and in an innovative program for another part. Even assuming no duplication of material from one program to the other, interactive effects will need to be taken into account. If teachers in a standard program try to capitalize on the learning experiences of the innovative program, can all achievement gain be attributed to the innovation? Are children likely to have negative reactions to the standard part of the program in contrast to the innovative part? Ways must be found to take account of these possibilities so that neither too little nor too much gain is attributed to the innovative program per se.

Selection of Students and Teachers. In the selection of teachers and students to participate in programs, there are a number of considerations. There is often a tendency (because of practical considerations) to fill programs with volunteers--both students and teachers. The motivation of these people may not reflect that which is typical in the larger setting, and thus the program results may not be applicable in the large. For this reason, random selection is to be preferred. If a program appeals to most of its target population, then something close to random selection can be obtained. In selecting experimental programs, one criterion should be their ability to obtain a representative sample from their target populations.

In a transient population there is always a serious problem of program dropouts. For evaluation results to be valid reflectors of long-term effects, a sufficient number of participants must be observed over a fairly long period of time. One alternative is to consider dropouts when determining initial sample sizes, but this is uncertain and costly. Another alternative is to select non-transient participants, but this may bias the sample. There is no easy solution, but its danger should be recognized.

III. SELECTION OF PROGRAMS

Concurrent with the planning for innovative programs should be the planning for future implementation. Since reproducibility of a program is always an implicit and often an explicit goal of an innovation, it should be one of the criteria used in the selection of programs. Many things can be done on a small scale that do not lend themselves to replication on a large scale. Class size can serve as an example. If a program is based upon the theory that in the primary grades achievement is best facilitated by instruction in very small groups, then a series of questions needs to be asked. Can the school district considering the innovative program afford the salaries for a sufficient number of teachers to implement it? Can the school district attract a sufficient number of teachers to carry on such a program? Does the school district have sufficient facilities to house such a program or would it need to build additional facilities, and if so, what is the likelihood that the funds to do so would be available? If all the answers are in the affirmative, then an innovative program based on the theory of small-group instruction would be a good idea. If, on the other hand, the answers are in the negative, alternative ways of meeting the objective might be explored. Could one teacher and two aides

carry on the program as well as could three teachers? If so, it might become feasible for most districts, in terms of personnel and facilities, and should be tried. If, on the other hand, the program can only be carried out by three teachers, each in a separate classroom, and could not be implemented by most districts, it probably should not be undertaken.

A host of problems relating to facilities, personnel, equipment and logistics need to be considered in the light of the desirability that innovative programs be reproducible. In the long run, it is better to grapple with them at the same time that decisions are made about introducing innovative programs than to be faced with the problem of trying to adapt a successful program to fit within a new set of constraints. This is closely tied to the necessity to identify the resource requirements of both the innovation and the operational program so that the cost may be determined.

IV. MEASURING ACHIEVEMENT AND INTERPRETING EVALUATIVE RESULTS

The specification of inputs and their organization into some sort of experimental design provides the framework for collecting observations on the alternatives being evaluated, observations that provide indicators of the relative effectiveness of those alternatives. It might be worthwhile to reflect on the question of what the observations can hope to show.

In the first place, the achievement-related variables that are measured will probably not be the real criteria on which one would wish to base his preferences; it is more likely that they will be surrogates that in some sense embody those criteria, but have the virtue of measureability.

Second, given the vagaries of attaching objective meaning to the utility concept in the context of education, evaluation can only go as far as making ordinal distinctions among alternatives. It is not clear, for example, that a program which produces a gain of .8 grade equivalents is worth twice as much as one which produces a gain of .4 grade equivalents, just as progress from the first to second grades probably does not have the same utility to society as progress between the third and fourth grades.

Accuracy. These observations notwithstanding, the usefulness of evaluative observations depends on their being accurate; this implies watchfulness in the selection of achievement-related variables, in the selection of instruments, and in the interpretation of test results. The objective is to provide basic data for a preference ordering with respect to effectiveness among competing program alternatives. Accordingly, the observations must have sufficient accuracy to enable the evaluator to discriminate consistently among those alternatives (i.e., if the experiment were replicated a number of times, the variance of the observations should be small enough so that comparisons among alternatives almost always lead to the same ordering of alternatives).

Achievement-Related Variables. The achievement-related variables obviously should reflect the educational goals of the research effort. They should also be measurable and relatively convenient to observe. However, it is also necessary to avoid the trap of measuring only that which is readily quantifiable. There is always a tendency to evaluate a program solely in terms of achievement gain because pre- and post-test measures are available in the form of standardized achievement tests. If there are other program goals, they must be evaluated even if measures need to be constructed for a specific purpose. If another program goal, for example, is to increase the ability of students to work together in the solution of a problem, this must be evaluated along with achievement gain.

Instruments of Measurement. The measuring instruments should reflect the instructional content of the programs being evaluated. If a given instrument provides observation on more than one achievement-related variable, and if the variables are eventually combined to compare against some composite criterion, then the testing instrument should reflect a balance consistent with the weights that are implicit in that criterion. The instruments must also be sensitive to differences in emphasis and timing of components of instruction. This is a critical factor when the interval between pre- and post-testing is short, and is in fact an argument against interim testing; after all, aside from social and political considerations, long-run educational objectives are more important than the short-run aspects of getting there. On the other hand, time is a luxury, and the interests of an efficient research effort encourage relatively frequent monitoring of the programs to keep them moving in profitable directions (such a pity that educational research lacks the speedy experimental vehicle that genetics found in the fruit fly). To suggest that instruments of measurement be sensitive to program differences is not to recommend that they be completely tailored to the differences; the objectivity of evaluation requires that measures be comparable despite program differences.

Scoring Modes. Finally, in interpreting measurements the evaluator must discern which scoring mode is appropriate for which purpose. Raw scores and grade equivalents are essentially absolute scores; percentiles and standard scores reflect achievement levels relative to those of all test participants. Because the distribution of absolute scores is wider for the higher grades, it is possible that over a period of years, a poor student may advance in terms of relative score (e.g., from the 20th to 25th percentile) while he is losing ground in terms of grade equivalents

gained; the question is whether the instructional program for that student is doing a more or less effective job than the instructional program of the average student who remains at the 50th percentile, while picking up a full grade equivalent each year.

Need for New Approaches to Measurement. It may be advisable to allocate a portion of research effort toward the study of relatively new measurement methods or the development of tools that are currently in the experimental stages. Sole reliance on traditional measures may reveal only part of what a program has to offer; it should not be surprising that innovative instructional programs may require innovative evaluation techniques.

BIBLIOGRAPHY

1. See recent studies by James S. Coleman, et al., *Equality of Educational Opportunity*, U.S. Department of Health, Education, and Welfare (1966); Jesse Burkhead, et al., *Input and Output in Large-City High Schools* (Syracuse, N.Y.: Syracuse University Press, 1967); Eric Hanushek, *The Education of Negroes and Whites*, unpublished doctoral dissertation (Department of Economics, Massachusetts Institute of Technology, 1968). For insights into the educational production process see Samuel S. Bowles, *Educational Production Functions*, in Education and Income. W. Lee Hansen (ed.) (New York: National Bureau of Economic Research, forthcoming); and Samuel S. Bowles and Henry M. Levin, *The Determinants of Scholastic Achievement*, The Journal of Human Resources (Winter 1968), pp. 3-24.

2. Carpenter, Margaret B., and S. A. Haggart, *Cost Effectiveness Analysis for Educational Planning*, The RAND Corporation, P-4327, March 1970.

3. Rapp, Marjorie L., *Evaluation as Feedback in the Program Development Cycle*, The RAND Corporation, P-4066, April 1969.