DOCUMENT RESUME

ED 030 208 EA 002 349

By-Marco, Gary L.

A Design for Evaluating Educational Programs in a Large City.

Pub Date Feb 69

Note-17p., Paper presented at the Annual Meeting of the Amer. Educ. Res. Assn. (Los Angeles, Calif., Feb. 1969).

EDRS Price MF-$0.25 HC-$0.95

Descriptors-*Educational Programs, *Measurement Techniques, Post Testing, Pretesting, *Program Evaluation, Statistical Analysis, Testing Problems, *Testing Programs

A pretest-posttest design for measuring the effects of educational programs uses comparison groups consisting of pupils like those in the treatment group but not getting that particular treatment. Although the design is geared primarily to evaluation of Title I programs in large cities, it should also apply to other situations. The plan for measuring treatment effects is to (1) identify the major objectives and the most important side effects of the programs, and (2) develop measures of the objectives and side effects. When several programs are being evaluated, careful selection of the comparison group is necessary to avoid the statistical problems of confounding and interaction of treatment effects. Evaluation of treatment effects is the process of judging the value of a treatment. In education, where the consequences of programs have much greater importance than the programs themselves, program evaluation is consideration of the consequences of the program. Therefore, program ratings should be based on how favorable and important the consequences are to the users. (LN)

A DESIGN FOR EVALUATING EDUCATIONAL PROGRAMS IN A LARGE CITY

Gary L. Marco

Educational Testing Service

In John L. Hayman, Jr. (Chm.), Management Problems in Conducting Large-Scale Research and Evaluation Studies. Symposium presented at the American Educational Research Association, Los Angeles, February 1969.

A DESIGN FOR EVALUATING EDUCATIONAL PROGRAMS IN A LARGE CITY

In education, program management is often restricted to the planning, the implementation, and later the replanning of programs. However, as I see it, the program management cycle can be broken down into seven steps:

1. Identifying the needs
2. Stating the program objectives
3. Selecting or developing criterion measures
4. Planning the program
5. Implementing the program
6. Measuring the effects
7. Evaluating the effects

After the last step, the cycle begins again with the reidentification of needs.

For each of these steps certain management problems could be identified. This morning I will focus on problems associated with measuring and evaluating the effects. Problems of measuring the effects are management control problems and thus fall primarily under the jurisdiction of the Director of Design and Test Development in Norton's plan (1969). However, I argue that evaluation problems are primarily those of stratagic planning and thus fall under the jurisdiction of the overall director. The word "evaluation" is often used to mean the process of measuring the effects of a treatment. I reserve the term to mean the process of judging the value of the effects of a treatment, whether with reference to cost or some other standard.

This morning, I shall first present a fairly simple design for measuring the effects of educational programs in a large city setting. Then, I shall talk about a major problem encountered in applying the design that program managers should be aware of. Finally, I shall discuss factors that a program manager should consider in evaluating the effects of a set of programs. The context I have in mind is the situation where (a) a variety of educational programs are represented and (b) the relative effectiveness of the programs must be evaluated. Although my remarks are geared primarily to the evaluation of Title I programs in a large city, they should apply also to other situations.

Measuring the Treatment of Effects

The design proposed for measuring the effects of a set of programs is a pretest-posttest comparison group design. A battery of tests (pretests) is administered when the pupil begins his involvement in a program; and the same or a different test battery (posttests), at the end of the involvement. Tests in the batteries are measures of the major objectives of the programs being evaluated plus measures of possible important side effects. Program participants take the complete batteries, even though some of the measures might be irrelevant to some of the objectives of any one program.

Identification of a Comparison Group

The design appears at this point to be a one group pretest-posttest design. However, where several programs with different objectives are being evaluated with the same measures, participants in one program can be used as a comparison group for participants in another program.

The key feature of a set of programs that permits the use of a comparison group is that not all of the objectives are relevant to all of the programs.

The need for a comparison group is clear to most of you. Only in those special situations where it can be assumed that nothing except the treatment produces the effect is a comparison group unnecessary. If an effect is produced, ideally one should not be able to point to anything other than the treatment as the cause of the effect. The purpose of using comparison groups is to rule out rival hypotheses explaining the effects alleged to have been produced by the treatment.

Campbell and Stanley (1963) have pointed to a number of rival hypotheses where a one group pretest-posttest design is employed, including history, maturation, the effects of testing, statistical regression, and selection. They recommended the pretest-posttest control group design as an alternative, since this design rules out a number of these rival hypotheses.

Unfortunately, in most school settings the pretest-posttest control group design cannot be used, for pupils cannot be assigned randomly to treatments without denying some needy pupils the benefit of the treatment. It would be unfair, for example, to deny remedial reading to half of the needy pupils just to satisfy the rigors of experimental design.

Even though a control group cannot always be identified for purposes of measurement, all hope for effective measurement is not lost. A comparison group other than a control group can be used. But the further the comparison group departs from the ideal control group, the more difficult it becomes to rule out rival hypotheses. In the

proposed design the comparison groups would consist of pupils like those in the treatment group but not getting that particular treatment.

The Plan

The plan I am proposing for measuring the effects of Title I programs would enable data on comparison groups to be gathered routinely. Pretests and posttests of the major objectives and important side effects of the programs would be administered to pupils in each of the programs. Thus, information on a particular objective would be available, not only for pupils participating in programs having as a primary goal that objective, but also for pupils in programs for which the objective is less relevant or irrelevant. (Of course, pre- and posttests could also be administered to selected pupils not participating in the programs of interest; that is, an external comparison group.) The pupils in programs designed to accomplish a certain objective would be expected to show more progress toward the objective than pupils in the other programs. If this result were not borne out in fact, the evidence would suggest that the program was ineffective in realizing that particular objective. In the case of reading programs, for example, scores of program participants would be compared with scores of participants in programs for which improved reading was not a primary objective.

As you have noted, the plan for measuring treatment effects consists of:

1. Identifying the major objectives and the most important side effects of the programs.

2. Selecting or developing measures of the objectives and side effects. (Conventional achievement batteries plus measures of school attitude and self-image would often be appropriate.)
3. Administering the test batteries to participants in the programs and possibly to a group of non-participants at the beginning and end of the treatment.

Since the design is intended for use in measuring general treatment effects, no measures of pupil background characteristics are included. When pretest and posttest measures are the same, measures of background characteristics are not useful as control variables, for the pretests are very efficient. However, they are useful as moderator variables and should be included in the pretest battery if the treatment could reasonably be expected to have differential effects on pupils in the treatment group.

I haven't as yet said anything that is very new to most of you. Perhaps I can rectify that situation as we look at programs associated with the application of the design.

Problems in Applying the Design

One problem is the interaction of selection and the treatment. The treatment effect may be moderated by the particular characteristics associated with the treatment group. In such a case, if the groups serving as the comparison and the treatment groups were interchanged, the effect of the treatment would not be the same; that is, there would be an interaction effect. I mention this factor only in

passing, for I have no solutions to propose at this time. The ideas mentioned in Trismen's paper (1969) may have some merit here.

Perhaps the biggest problem in the proposed design is the confounding of treatment effects. Many pupils often participate in several programs. This is particularly true in the case of programs for disadvantaged pupils. Several treatments thus affect each pupil, so that it is exceedingly difficult to isolate the effects of a single treatment.

A simplified description of treatment overlap is shown in Table 1, which is in the Appendix. It is assumed that the treatments are supplemental to regular instructional programs. Four enrollment patterns are readily identifiable when one considers all related programs and all unrelated programs, respectively, as single units. Related programs are those programs which have as one of their major objectives the objective under consideration.

As you can see on the second page of the Appendix, the pretest-posttest difference score of an individual in the treatment group may be expressed as follows:

$$\text{Difference Score} = \text{Program Effect} + \text{Effect of Related Programs} + \text{Effect of Unrelated Programs} + \text{Effect of School} + \text{Individual Effect}$$

It is assumed that there are no interaction effects among treatments. Under the same additive model the mean difference score for a particular treatment group will be equal to:

$$\text{Mean Difference Score} = \text{Mean Program Effect} + \text{Mean School Effect} + (\%\text{ of Group in Related Programs} \times \text{Mean Effect of Related Programs}) + (\%\text{ of Group in Unrelated Programs} \times \text{Mean Effect of Unrelated Programs})$$

You may note that the individual effects are assumed to have a mean of zero and cancel out.

It seems reasonable to assume also that unrelated programs have a negligible effect on the pretest-posttest difference score. Thus, the last component can be eliminated. If, in addition, all other outside influences, including the effects of related programs and school, have a constant effect from program to program, there is no need to consider them in estimating program effectiveness; for since one is interested in the difference between programs, the constant effect cancels out. Thus, the mean difference score for a particular treatment group would then be equal to:

Mean Difference Score = Mean Program Effect + Constant Effect of Outside Influences

In this special case, there is no need for comparison groups--unless one is interested in the absolute rather than the relative effects of the programs.

Unfortunately, the proportion of pupils participating in related programs may vary considerably from program to program. Thus, even if the effect of outside influences on such programs is constant (or nearly so), the contribution to the difference score for a particular treatment group may vary considerably as a result of unequal participation in related programs. Utilization of a comparison group not in the program but having similar proportions enrolled in related programs allows one to eliminate the effects of related programs and school. The problem is that one must find a comparison group that has the same pattern of enrollment in related programs as the treatment

group. It is difficult to keep track of the enrollment patterns of pupils, but matching them is an even more difficult problem.

One alternative is to use as a comparison group pupils not enrolled in the program of interest nor in related programs, and to estimate the effect for those pupils in the treatment groups also not enrolled in any related programs. Unfortunately, since multiple enrollment is often the rule, the sizes of the treatment and comparison groups identified in this fashion might be quite small.

A possible solution to the problem is suggested by profile analysis. The profile for a given program is taken to be the proportion (p) of its pupils enrolled in each specific treatment. A complete profile matrix is shown in Table 2 in the Appendix. You will note that 1's are in the diagonal, indicating that all of Group 1 is enrolled in Program 1, etc. However, the matrix is non-symmetric. Thus, the proportion of Group 1 enrolled in Program 2 does not necessarily equal the proportion of Group 2 enrolled in Program 1.

An appropriate comparison group for Group 1, the treatment group for Program 1, would be a group with a p in column 1 of zero (that is, one not participating in Program 1) and with the remaining p's matching the p's for Group 1. Of course, since one must have a comparison group with a p of 0 for Program 1, the pupils in the other programs who are enrolled in Program 1 must be eliminated. The non-treatment groups are thus redefined, and the profiles are recalculated. The group having a profile on related programs most like

the profile of Group 1 according to a distance measure of profile similarity would be used as the comparison group.

A better procedure would be to combine groups to provide a profile similar to that of the treatment group. Unfortunately, the number of combinations that one would have to look at is burdensome unless a computer is used or the number of programs is small. For example, the number of combinations one would have in the case of eight possible comparison groups would be the sum of the combination of eight things taken one at a time all the way up to the sum of eight things taken eight at a time, or 255 combinations! For a "combination" comparison group to be identified, it would be necessary to redefine the groups, not only to eliminate overlap with the treatment group, but also to eliminate overlap with other groups. One way to do this is to eliminate those pupils from Group 2 who are also enrolled in the treatment group, say Group 1; those pupils from Group 3 who are also enrolled in the treatment group or in Group 2; etc.

A procedure that requires no group redefinition is the formation of a comparison group from a combination of pupils who are not in the treatment group. However, the process of finding a combination of pupils that has a profile on related programs like that of the treatment group would be prohibitive in the usual situations where a large number of pupils is involved.

Evaluating the Treatment Effects

The Need to Evaluate

So far I have presented a design for measuring treatment effects and discussed a major problem in implementing the design. But program managers encounter difficulties in evaluating as well as in measuring the effects of programs. It is the evaluation process that I wish to examine now.

I call your attention to the definition of evaluation mentioned earlier; namely, the process of judging the value of a treatment. Often in the case of a single program, the treatment effects are given criterion by criterion, and no attempt is made to evaluate the treatment in the sense just mentioned.

However, as John Dewey (1916) made eminently clear in his discussion of judgments of value, judging the value of, or evaluating, a treatment is necessary if a choice is to be made among alternative courses of action. Evaluation is necessary to make that decision, but it is not necessary if no alternatives exist; for example, if a program would continue unchanged no matter what the effects of the program might be. Since the purpose of comparing the effects of a variety of programs is usually to decide how to allocate resources, the need for evaluation should be clear.

How to Evaluate

As you can see from Table _ in the Appendix, the value judgments required in evaluation may be either an overall evaluation of each program or several evaluations--one for each of the major objectives,

or criteria, being considered. A cell entry in the table refers to the value of a particular program in accomplishing a specific objective. An entry at the end of a row refers to the general evaluation of a particular program. Such data can be used as input to a decision model like that to be described in Badran's paper (1969).

The problem is how to arrive at an index of value, whether it be a cell entry or an overall evaluation. Educational programs are not usually considered to be of intrinsic value; it is their consequences that are of interest. Thus, program evaluation is nothing more nor less than consideration of the consequences of the program. The evaluation process is an inductive process that takes the following form:

Program x causes consequence y.

Consequence y is good (bad).

Therefore, x is good (bad).

To evaluate a program one must first determine the consequences and then evaluate them.

Consider the evaluation of a program--call it Program A. Suppose that Program A is reported to have had the following consequences:

1. The average gain score on Test A over a nine-month period was 10 points compared to 3 points for a comparison group.
2. Three favorable and five unfavorable newspaper reports have been printed during the year.
3. Seventy-five percent of the teachers involved in the program desire to teach in the program next year.

4. The average daily attendance of pupils enrolled in the program is 155 days (out of 180) compared with 165 days for pupils in a comparison group.

It should be recognized that these consequences may or may not be true. Moreover, some of the consequences are more favorable; and some, more important than others.

The task of the judge, or evaluator, is to weigh the evidence in order to arrive at an evaluation of Program A. Weighing the evidence in this case is not unlike the task of a petit jury in a court of law. The jury must decide whether or not a particular consequence, or piece of evidence, is true, how favorable it is, and how important it is. Then he must arrive at an overall judgment based on all of the evidence.

Either an analytical or a clinical procedure may be used to arrive at a value judgment. The analytical procedure would involve specific judgments on scales like those shown on the second page of the Appendix. An overall rating could be computed by multiplying the scores on the separate scales.

The dimension included here that conventional rating scales do not include is the importance, or relevance, dimension. The dimension is necessary if evidence is to be weighted differentially. Another feature of evaluation that is not common to most rating procedures is that the scale values on all three scales are allowed to vary from judge to judge. Fishbein (1967) developed a rating procedure that allowed ratings to vary in this manner--but only on the true-false and favorable-unfavorable scales.

It is on the basis of the truth, the favorableness, and the importance of the program consequences that a program must be evaluated. However, in order to be convincing, the claims to truth or falsity, favorableness or unfavorableness, and importance or unimportance must be justified to the audience. I do not have time to elaborate on the process of justification. Aschner (1956) saw reasons and rules as important to justification. Let me say simply that good reasons must be given--good in the sense that they are acceptable to the audience.

Recommendations

So far this morning I have talked about measuring and evaluating treatment effects when several treatments are administered. I would like to close by making these recommendations to program managers.

1. In measuring treatment effects for a variety of programs, consider using the participants in programs unrelated to the treatment as a comparison group.
2. In instances where program overlap is great, choose a comparison group that is involved in the same type of related programs as the treatment group.
3. When alternative courses of action are being considered, make sure you consciously evaluate in the true sense of the word the effects of the courses of action.
4. In evaluating a program, determine the consequences of the program and assign your rating on the basis of how <u>favorable</u> and <u>important</u> the consequences are to <u>you</u>.

5. If you have other people evaluate programs for you, make sure you understand the basis on which they make their evaluations.

References

Aschner, M. J. Teaching the anatomy of criticism. School Review, 1956, 64, 317-322.

Badran, Y. I. A model for making decisions about educational programs. In John L. Hayman, Jr. (Chm.), Management Problems in Conducting Large-Scale Research and Evaluation Studies. Symposium presented at the American Educational Research Association, Los Angeles, February 1969.

Campbell, D. T., & 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.

Dewey, J. Essays in experimental logic. (Dover Publications ed.) Chicago: University of Chicago Press, 1916.

Fishbein, M. A consideration of beliefs, and their role in attitude measurement. In M. Fishbein (Ed.), Readings in attitude theory and measurement. New York: Wiley, 1967, Pp. 257-266.

Norton, D. P. Management emphases in the installation of an evaluation staff for federally-funded projects in a large city. In John L. Hayman, Jr. (Chm.), Management Problems in Conducting Large-Scale Research and Evaluation Studies. Symposium presented at the American Educational Research Association, Los Angeles, February 1969.

Trismen, D. A. Replication: A substitute for replicability. In John L. Hayman, Jr. (Chm.), Management Problems in Conducting Large-Scale Research and Evaluation Studies. Symposium presented at the American Educational Research Association, Los Angeles, February 1969.

Appendix

Table 1

Enrollment Patterns

	Enrolled in:			
	Particular Program	Related Programs	Unrelated Programs	School
Pattern A	X			X
Pattern B	X	X		X
Pattern C	X	X	X	X
Pattern D	X		X	X

Table 2

Profile Matrix

	Program			
	1	2	...	J
Group 1	1.0	p_{12}	...	p_{1J}
Group 2	p_{21}	1.0	...	p_{2J}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Group I	p_{I1}	p_{12}	...	1.0

Table 3

Indices of Value

	Objective				
Program	1	2	...	K	Overall
1	V_{11}	V_{12}	...	V_{1K}	$V_{1.}$
2	V_{21}	V_{22}	...	V_{2K}	$V_{2.}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
J	V_{J1}	V_{J2}	...	V_{JK}	$V_{J.}$

Difference Score Models

Assuming additivity:

Difference Score = Program Effect + Effect of Related Programs + Effect of Unrelated Programs + Effect of School + Individual Effect

Assuming additivity:

Mean Difference Score = Mean Program Effect + Mean School Effect + (% of Group in Related Programs x Mean Effect of Related Programs) + (% of Group in Unrelated Programs x Mean Effect of Unrelated Programs)

Assuming additivity and constant effect of outside influences:

Mean Difference Score = Mean Program Effect + Constant Effect of Outside Influences

Judgments-of-Value Scales

True		1		0		False
Favorable	2	1	0	-1	-2	Unfavorable
Important	4	3	2	1	0	Unimportant