DOCUMENT RESUME

ED 064 412 TM 001 691

AUTHOR Tatsuoka, Maurice M.
TITLE Nationwide Evaluation and Experimental Design.
PUB DATE Apr 72
NOTE 16p.; Paper presented at the annual meeting of the AERA at a symposium on "Some Problems Associated with Nationwide Evaluation of Educational Problems in Schools," (Chicago, Ill., April 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Decision Making; *Evaluation Techniques; Improvement; *Research Methodology

ABSTRACT

A discussion of pay-off evaluation is presented. Various objections to experimental design are raised. These include: (1) the requirement of random assignment of units of analyses to treatment and control conditions, (2) the "conflict with the principle that evaluation should facilitate the continual improvement of a program, and (3) the fact that evaluation is almost useless as a device for making decisions during the planning and implementation of a project. (CK)

NATIONWIDE EVALUATION AND EXPERIMENTAL DESIGN

Maurice M. Tatsuoka

University of Illinois at Urbana-Champaign

Prepared for

a symposium on "Some Problems Associated with Nationwide Evaluation of Educational Problems in Schools," 1972 Annual Meeting of the American Educational Research Association

NATIONWIDE EVALUATION AND EXPERIMENTAL DESIGN

Maurice M. Tatsuoka
University of Illinois at Urbana-Champaign

In the introduction to his now-classic paper on "The Methodology of Evaluation" (1967), Professor Scriven remarked that when a newcomer stands on the shoulders of giants to see farther in a given field, "this feat is often confused with treading on their toes." With Scriven it was unmistakably a case of one giant standing on the shoulders of other giants, for he has provided us with many new vistas and insights. What I am about to undertake now, however, will, I regret to say, far more closely resemble the act of treading on the giants' toes. Nevertheless, I would like to believe that this act is not quite so obnoxious as it may first seem, and is even not entirely devoid of merit. To have their toes trodden upon by a dwarf like me cannot possibly hurt the giants--unless, of course, they have corns on their toes. In the latter event they will, hopefully, take measures to remove their corns.

In this paper, I shall confine my attention to pay-off evaluation, not because I consider intrinsic evaluation any less important, but because the former is the only aspect for which experimental design is relevant. Many evaluators would, I realize, say that experimental design is not relevant to any type of evaluation. I shall do my best to blast this view, which has perhaps been most explicitly and eloquently stated by Stufflebeam (1969).

The first objection to experimental design raised by Stufflebeam concerns the very requirement of random assignment of units of analyses to treatment and control conditions, which is claimed to be all but impossible in the context of evaluation studies. But he supports this contention only by citing the case of random assignment of individual students to conditions. Surely we all agree that individual students are not the appropriate units in large-

scale program evaluation. Classes, schools, or even school districts are the proper units, and random assignment of these to the conditions is not nearly so infeasible as that of students.

Of course I realize that, even with these larger units of analyses, random assignment is not so simple as in laboratory experiments. There are many administrative, logistic, and political problems to be solved before random assignment can be achieved. Partly because of these problems, Stake (1969) has proclaimed the "need for limits" in program evaluation. I shall return to this point later, but first let me continue treading on Professor Stufflebeam's toes. Besides the individual versus larger-unit distinction, he fails to acknowledge that, as a last resort, we can give "deferred preferential treatment" to those units that happen to be assigned to the control or "non-treatment" condition in order to overcome administrative resistance to exclusion from a presumably beneficial program. That is, we can promise (and of course honor our promise) that the units assigned to the control condition will be given the experimental treatment in the following year. (Of course, as Glass (1971) points out, we would thereby sacrifice the "opportunity for long-range comparison of groups." But this seems to be a minor loss compared to the preclusion of random assignment.)

It has also been objected (although not by Stufflebeam in the paper I'm now referring to) that random assignment is unethical, for we would be depriving some units (i.e., classes, schools, or school districts) of the benefits of the new program. This argument would be pertinent only if it were known a priori that the new program is indeed beneficial (in which case there would be no need for an evaluation) and if funds were available for implementing this program across the board throughout the nation. Since few if

any newly proposed programs can satisfy both these conditions, the argument of "unethicality" loses most of its force. When at least one of these conditions is not met, what could be more ethical (and democratic, if you please) than a completely random assignment to treatment and non-treatment?

Stufflebeam's second objection to experimental design stems from its alleged "conflict with the principle that evaluation should facilitate the continual improvement of a program [p. 49]." His basis for this contention is the belief that experimental design requires us to hold the treatment constant throughout the experiment, thus stifling dynamic development of programs based on continual feedback of how they are working. Such a belief seems to me to reflect a misconception of what constitutes a treatment in the program-evaluation context. True, in a laboratory experiment in which the treatments are completely specified a priori--such as fixed dosages of a drug, or certain methods of stimulus presentation--these must be held constant throughout. But an educational program is, by its very nature, an entity that is in perpetual flux. Only some broad guidelines and principles are typically specified at the outset, and details of how to carry out the program are usually left to the individual administrator to plan and modify with experience. This fluid, dynamic entity, with all its periodic modifications and refinements IS the treatment. Nothing in experimental design forbids such types of treatment. All that is required is that an accurate running record be kept of what sorts of modifications and refinements were made at what stage for what reasons, so that upon completion of the evaluation we can describe what it is that has been evaluated.

The next indictment against "the experimental design type of evaluation" made by Stufflebeam is that "it is useful for making decisions after a project

has run full cycle but almost useless as a device for making decisions during the planning and implementation of a project [p. 49]." This is little more than a rephrasing of the second point discussed above. It is, of course, trivially true that the final or summative evaluation results cannot help in making intermediate decisions; only formative evaluation is relevant for these. But the summative-formative distinction (Scriven, 1967) refers to two different roles of evaluation, and not to the different methodological types of evaluation such as experimental-design or non-experimental-design types. It seems to me that using an experimental design in no way forbids the intermediate monitoring of feedback information, nor--for reasons discussed above--does it forbid acting on such information for periodic modification and refinement of the program. If anything, it should enhance the generality of the information thus monitored (or at least that part of the information based on inter-program comparisons), because of the random-assignment base.

The pointing out of the next alleged flaw is attributed to Guba (1965), that "experimental design is well suited to the antiseptic conditions of the laboratory but not to the septic conditions of the classroom [p. 50]." In elaboration it is asserted that, in order to apply an experimental design, "the potential confounding variables must be either controlled or eliminated through randomization" [emphases added]. Surely this is not the case. The confounding variables, if clearly identifiable and sufficiently important, may be used as stratifying or "blocking" variables in a factorial design--rather than being "controlled" in the narrow sense of being fixed and prevented from operating as variables. Randomizing doesn't eliminate them, but assures us that, in the long run, they will be uncorrelated with the treatment variables. (And this is why randomization is so important.) Thus,

there is no reason why experimental design cannot be used under the "septic conditions of the classroom." Of course, the observed effects will not be completely attributable to the treatment variables, but will have to be allocated partly (sometimes even mostly) to the "confounding variables." But this is due to the nature of things in most real-life situations, and not to the use of experimental design. On the contrary, using an experimental design enables us to estimate what percentage of the observed differences may be attributed to the treatment and what percentage to the "confounding variables." Far from muddying up things, it achieves whatever measure of clarification and ordering that is possible under the circumstances.

Finally, Stufflebeam claims that "while internal validity may be gained through the control of extraneous variables, [this] is accomplished at the expense of external validity [p. 51]." This contention is again based on a narrow conception of what is meant by "controlling" an extraneous variable. As pointed out above, control need not take the form of actually fixing the variables. Only when relevant extra-treatment variables are controlled in this sense will generalizability to the real world be sacrificed. Such an eventuality is not engendered by experimental design as such, but by an inexpert use of it.

I now come back to the deliberate limitation on generalizability advocated by Stake (1969), alluded to earlier. His reason for so advocating is that he believes the two questions, "What is at work in the program?" and "Why does it work?", cannot be simultaneously answered by a single type of study. To find out why, he says, a strict, laboratory-type controlled experiment must be done, in which case "the program being researched [often] no longer is the program [we] wanted to know about [p. 40]." To investigate what, he

continues, we need descriptive and judgmental evaluation studies of limited generalizability. Ergo, evaluation studies (at least in their summative role) should be concerned only with the what question with regard to a specific program in a specific setting and should forget about generalizability to other settings.

It seems to me that the above argument contains several flaws. Certainly, a "pure science" type of experiment--in which rigid controls are exercised and systematic variations are introduced in accordance with a pre-planned schedule--will generate a "test-tube program" bearing little resemblance to what may be expected to operate in real-life settings. (I'm a bit puzzled why Stake recommends this type of study for formative evaluation done for the benefit of the program developers and for "broadcast[ing] to a wide audience of educators and researchers" who want to know if the program will work in other settings--since that which will generalize would be a test-tube program, not a real one.) But surely there must be at least two subclasses of why questions: those that admit of answers only by recourse to lab-type experiments, and those that are answerable by use of experimental designs in which the fluid, dynamic entities that real-life programs are, constitute the treatments. We might label these, in Carnapian style, the why_1 and why_2 questions, respectively. I suspect that Stake had the why_1 questions in mind when he warned that answering why questions would alter the program, but was thinking of why_2 questions when he said that formative evaluation studies would address themselves to why questions.

To simplify the notation, let me hereafter drop the subscript 1 in "why_1 questions" and refer to the why_2 questions as "how questions." So we now have why, how, and what questions, in descending order of "basicness." The

why questions are in the province of basic instructional research--which should eventually provide us with general principles that will help us in planning and developing educational programs--but they are not of immediate concern to present evaluation endeavors. The how questions ask what sorts of components (or "transactions," to use Stake's (1967) terminology) in a program are associated with what kinds of outcome under different antecedent conditions--and, hence, "How can we replicate these outcomes in another setting?" Answers to these questions clearly need to be generalizable in order to serve any purpose. And I contend that at least tentative answers can be obtained without doing lab-type experiments, but using experimental design in a liberalized sense, as outlined below. I think it was this kind of research that Hastings (1966) had in mind when he called upon evaluators to pay more attention to "the why of the outcomes."

Thus, I believe that it was an unduly narrow construction of "why questions" that led Stake to hold what seems to me an untenable position that evaluation should be concerned only with what questions, yielding specific and ungeneralizable answers. Another reason why such a position is untenable was given by Wardrop (1969), who pointed out that, "whether or not an evaluation study is designed for generalizability, the consumer will make generalizations from its results [p. 41]." Thus the evaluator has a moral obligation to design his study for maximum generalizability within the constraints under which he operates.

I have concluded my toe-treading act. It is now time to offer my own penny's worth of ideas and permit the giants to trample me down if they wish. But before that, let me anticipate one possible reaction which many evaluators may have to my foregoing remarks. "Okay. So you've stretched

the concept of experimental design to allow modifying the 'treatment' in mid-stream," the reaction might run, "but then you're no longer talking about the kind of experimental design that we're objecting to. All you've done is to pull a semantic sleight of hand." In a way, this may be true; but not completely. In objecting to experimental design, many evaluators seem to be rejecting the essential principle of random assignment of units to treatment conditions, besides the lab-oriented principle of constancy of treatment throughout the experiment. (Recall Stufflebeam's explicit statement to this effect and Stake's advocacy of deliberate limitation of generalizability.) My continuing to use the term "experimental design" in a "stretched-out" sense (I'd prefer to call it a liberalized sense) thus serves, if nothing else, as a preventive measure against throwing the baby out with the bath water: constancy of treatment may--and, in the evaluation context, should--be thrown out; randomization must not.[1]

Enough procrastination! I now stick my neck out. The way I would go about the gigantic task of evaluating a nationwide intervention program such

---

[1] I realize that, in some cases, the political obstacles against random assignment are simply insurmountable. Title I of the 1965 Elementary and Secondary Education Act, for instance, required that all eligible school districts (i.e., those with a given percentage or more of disadvantaged children) be included in the program. In such cases, it seems to me that there are only two alternatives available. Either evaluators as a group must turn to politics and lobby (or otherwise seek to modify the political climate) for bringing about a change in the law, or we must resort to a quasi-experimental design such as the interrupted time-series design, in which the past history of each experimental unit serves as its own "control group." Since, as Cohen (1970) points out, the evaluation of a nationwide intervention program is in any case partly a political activity, the first alternative is not so outlandish as it may first seem. However, since the change of laws is a time-consuming process, we will probably have to adopt the second alternative while we are waiting. Quasi-experimental designs suitable for evaluating social intervention programs have recently been discussed at length by Campbell (1969), who gives interesting examples of actual evaluation studies using these designs.

as Headstart, Title I, or Followthrough, would be somewhat as follows. (Remember that I'm dealing only with the pay-off evaluation aspect in this paper.)

(1) Assign a large number of instructional units (classes, schools, or school districts) at random--or, more realistically, on a stratified random basis--to the experimental and control conditions, invoking the "deferred preferential treatment" clause if necessary.

(2) Obtain descriptive data on antecedent conditions for each unit in as great a detail as possible.

(3) Specify only broad guidelines of the program for administrators of the experimental units. Leave details, modifications, and refinements up to the individual administrator, to be made in his best judgment as experience accumulates. Require accurate chronological recording of specific transactions.

(4) After the program has run one full cycle, obtain measures on whatever variables are related to the general objectives of the program--be they cognitive, affective, or conative--using comparable instruments across the nation. (This is not to say that intermediate measures should not be taken during the course of program implementation for program-modification purposes. However, only the final measures will be used in the analyses.)

(5) Carry out a multivariate analysis of variance of the data obtained in (4), using the bases of stratification (if any) adopted in (1) as additional factors besides the main treatment factor (experimental vs. control). Suitable covariates, such as average IQ of students in the instructional unit, may be used if these have not already been used as stratifying variables.

Up to this point, the strategy I'm proposing is superficially similar to those proposed by Light and Smith (1970) and by Glass (1971). But there is one major difference. Both these papers permit only pre-planned variations of the program. Light and Smith introduced pre-planned variations to overcome the defect they saw in the Westinghouse-Ohio analysis (Cicerelli et al., 1969) of the Headstart program: "that, except for the overall difference between the [experimental and control groups], all the differences in performance between the two groups, from town to town, was attributed to chance [p. 13]." Glass justified his proposal for pre-planning on the grounds that "probably no more than about six prototypical programs for disadvantaged pupils are required to capture the range of plausible intervention strategies [p. 8]." But this a priori specification of immutable treatments is precisely the reason why many evaluators reject experimental design. And it is my thesis that this aspect of lab-oriented experimental design is the bath water we should throw away, once we recognize that the entity we want to evaluate is the dynamic one of a program in flux, and not a program rigidly specified in advance.

The next phase of my proposed strategy is admittedly ex post facto, and I know that it is now the experts in experimental design, rather than evaluators, who would thumb me down. Campbell (1969) has "totally rejected" ex post facto designs "because of the specific methodological trap of regression artifacts [p. 411]," and Glass (1971) condemns any reliance on them as a pernicious habit that hinders the widespread acceptance of planned experimentation in evaluation circles. For reasons described below, I feel that their positions on this matter are too extreme.

In a nutshell, my proposal is to group the many spontaneously generated

variants of the program into such categories as "very good," "good," "fair," and "poor" in terms of their outcomes measures in step (4) above, and to investigate--perhaps by means of multiple discriminant analysis--the antecedent- and transaction-variable combinations that best differentiate the good from the poor variants of the program. This would allow us to generate hypotheses as to which versions (as described by the detailed record of transactions collected in step (3) above) are likely to work best under what sort of settings.

As soon as I let it be known that my proposed use of an ex post facto design is only for the purpose of generating hypotheses, Campbell and Glass, among others, are likely to say, "Oh, then it's okay. Why didn't you say so in the first place?" and they would probably accuse me of having erected straw men to attack when I commented that their position with respect to ex post facto designs was too extreme. They never did (they would retort) condemn these designs as tools for generating hypotheses for future, independent testing, but only as devices for drawing conclusions from data at hand. True, but nor did they explicitly mention that these designs could be useful in generating hypotheses--at least not in the papers referred to above. My point is that their extreme pronouncements (I'll withdraw the word "positions") could easily be misinterpreted as an across-the-board condemnation of ex post facto designs for any and all purposes.

In the interest of compactness, I glossed over several difficulties when describing the second phase of my proposed strategy above. I realize that it is no easy matter to form "good" to "poor" groups of the many variants of the program. Furthermore, the grouping should not be done solely on the basis of the pay-off analyses, but should include the results of intrinsic

evaluations as well. Some kind of weighting of the verdicts from the two types of evaluation will have to be made, as Scriven (1967) has indicated, and this is a difficult problem to solve. Perhaps, as Glass (undated) suggests, the ultimate answer may lie in the construction of a "fundamental scale of utility" for assessing the intrinsic properties and the outcomes of a program with a common yardstick. But this will probably not happen for many years to come. Alternatively, we could leave the various merit-criteria in multivariate form, and not group the competing program versions at all. We would then have three sets of variables describing, respectively, the antecedents, the transactions, and the merit indicanda (intrinsic properties and outcomes). A generalized canonical correlation analysis for more than two sets of variables, developed by Horst (1961), could then be used to analyze these data. In fact, this approach would be superior to the discriminant analysis suggested above, because it would keep the antecedent- and transaction-variables sets physically distinct from each other. It will thus be easier to generate hypotheses as to which program version under what setting would likely lead to best outcomes.

Then what? As you have probably guessed, we would launch a second cycle of the first phase of the proposed evaluation strategy, with somewhat more detailed specification of program guidelines in step (3) for the majority of instructional units, but the same broad guidelines as in the first cycle for the remaining few. The more detailed specifications would be based on the transaction descriptors of those variants that were judged, say, at least "fair" in the first cycle. The broad-guideline-only units are included in recognition of the fact that the "poor" variants were so judged in the first cycle only on the basis of an ex post facto analysis. If similar variants

are generated in the second cycle, they may prove to be not so poor.

After the program has run its second cycle, the second phase (the ex post facto phase) of evaluation would again be undertaken; and the cycle would repeat--but, hopefully, not ad infinitum! My expectation is that, with successive iteration cycles, fewer and fewer variants will remain to be evaluated, and these will become better and better for their respective kinds of setting. The program specifications will get tighter and tighter, but there will be less and less need for drastic departures from them, and "convergence" will, perhaps, eventually be achieved--until, of course, a new set of antecedent conditions emerges.

Is it utopian to expect such sustained evaluation efforts over an indefinite number of cycles? Perhaps so. But, given the fact that a program--especially a nationwide social intervention program--is a dynamic entity, and given a commitment to experimental design for maximal generalizability, I cannot see how the evaluation can possibly be a one-shot affair. Some kind of iterative cycling seems mandatory. In saying this, I am, of course, concurring with Professor Stufflebeam's idea of cycling and recycling, inherent in his CIPP (context-input-process-product) model--but with one difference. He does not seem to consider randomization to be terribly important, while I regard it as essential; in fact, a fresh randomization needs to be done for each cycle--even if only a restricted randomization may be possible under the constraint of giving preferential treatment (again on a random basis) to some of the units that were denied it in the preceding cycle. I am also agreeing with Professor Stake (1969) that, in a certain sense, the distinction between formative and summative evaluation is an academic one: a summative evaluation for one cycle is a formative evaluation for the next.

References

Campbell, D. T. Reforms as experiments. American Psychologist, 1969, 24, 409-428.

Cicerelli, V. G., et al. The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development. Washington, D. C.: Westinghouse Learning Corporation and Ohio University, 1969.

Cohen, D. K. Politics and research: The evaluation of large-scale social action programs in education. Review of Educational Research, 1970, 40, 213-238.

Glass, G. V. Pay-off evaluation of Title I. (Paper delivered at a symposium entitled "National Evaluation of Title I" 1971 annual convention of the National Council on Measurement in Education.) New York, February, 1971.

Glass, G. V. The growth of evaluation methodology. Boulder, Colorado: Laboratory of Educational Research, University of Colorado, undated. (Mimeo.)

Guba, E. G. Methodological strategies for educational change. (Paper delivered at the Conference on Strategies for Educational Change.) Washington, D. C., November, 1965.

Hastings, J. T. Curriculum evaluation: The why of outcomes. Journal of Educational Measurement, 1966, 3, 27-32.

Horst, P. Relations among m sets of measures. Psychometrika, 1961, 26, 129-149.

Light, R. J., & Smith, P. V. Choosing a future: Strategies for designing and evaluating new programs. Harvard Educational Review, 1970, 40, 1-28.

Scriven, M. The methodology of evaluation. AERA monograph series or curriculum evaluation, No. 1. Chicago: Rand McNally, 1967. Pp. 39-83.

Stake, R. L. The countenance of educational evaluation. Teacher's College Record, 1967, 68, 523-540.

Stake, R. L. Generalizability of program evaluation: The need for limits. Educational Product Report, 1969, 2, 39-40.

Stufflebeam, D. L. Evaluation as enlightenment for decision making. In Improving educational assessment and an inventory of measures of affective behavior. Washington, D. C.: Association for Supervision and Curriculum Development, National Education Association, 1969. Pp. 41-73.

Wardrop, J. L. Generalizability of program evaluation: The danger of limits. Educational Product Report, 1969, 2, 41-42.