EVALUATING OUTCOMES: SOME PROBLEMS OF INTERPRETATION*

J. Robert Warmbrod
Professor

Agricultural Education
The Ohio State University

"Our group had 20 per cent fewer (dental) cavities!" All of us have heard this or a similar report of the findings of studies designed to evaluate the effectiveness of toothpaste. When we attempt to interpret this outcome or finding, some questions immediately come to mind: With what other group or groups was "our group" compared? How were participants selected? What was the precise nature of the treatment? How old were the participants? What was the condition of their teeth before the experiment? What measurements were made? Who made the measurements?

We could make a long list of questions. The important point is that the answers to these questions have a great deal to do with how we interpret the results claimed for a particular brand of toothpaste.

A closer look at the questions reveals that some actually raise doubt about the truthfulness (validity) of the finding. How sure are we that a certain brand of toothpaste was the factor which really reduced the number of cavities? Perhaps there were factors operating, other than the toothpaste, which could help reduce the number of cavities. Each of these factors could contribute to a reduction in the number of cavities regardless of the brand of toothpaste used; or at best, when use of the toothpaste was accompanied by certain conditions, it may be more effective than when used under more ordinary conditions.

---

*This article is based on a paper presented during a Workshop on Developing Procedures and Techniques for Evaluating Manpower Development and Training Act Programs conducted by the Kentucky Research Coordinating Unit in Louisville, Kentucky, May 1970. A copy of the complete paper is available from the author.

Journal of the American Association of Teacher Educators in Agriculture
Volume 11, Number 2, pp.1-10
DOI: 10.5032/jaatea.1970.02001

The skeptic's questions call attention to another dimension of interpretation. Even if we are satisfied that the finding is legitimate -- that is, the brand of toothpaste made a difference in the experiment -- parents may still wonder whether similar results will be forthcoming if their children use the toothpaste. So valid outcomes in particular settings involving certain groups of people at a particular point in time are not automatically and completely applicable to what might be expected to occur in other settings with different groups of people at some other time.

I hope the example helps clarify and pinpoint two important questions that must be asked about evaluative studies designed to assess the outcomes of educational programs. Instead of the number of cavities, we are concerned with outcome measures such as change in behavior, ability to perform, and actual performance of persons who complete or leave vocational education programs. First, we need to assess the extent to which the outcomes observed can actually be attributed to the educational program -- Did the treatment (educational program) really make a difference? If we can rule out or discount some other factors which may have caused or contributed to the outcomes observed, then we place a different interpretation on the findings than would be the case otherwise. If our curiosity is satisfied on the first question, the second category of suspicions probes this question: To what groups, settings, situations, and types of measures do the results apply?

To What Extent Are the Findings Valid?

Some Preliminary Considerations

Description, Correlation, and Cause-Effect. In interpreting evaluations of educational programs, we need to pay close attention to the intent or purpose of the evaluation. Most evaluative studies can be placed in one of three categories which indicate a basic purpose or intent. The categories become, then, labels for claims made by the evaluator.

First there are studies designed primarily to describe the characteristics, competencies, or performance of persons who have completed a vocational education program. In descriptive studies the evaluator's primary claim (or intent) is to describe, in a complete and accurate manner hopefully, the characteristics,

kind and level of competence possessed, or level of performance of persons who have been exposed to a particular series of educational experiences necessarily produced or caused the outcomes observed. Unfortunately, the evaluator who sets out to describe outcomes, the consumer of the findings, or both frequently imply or claim a more direct relationship between program and outcomes than the design of the study allows or warrants.

A second category includes evaluative studies designed to investigate relationships between outcomes and characteristics of enrollees (e.g. age, socio-economic status), between outcomes and various dimensions of the instructional program (e.g. number of hours supervised occupational experience), or between outcomes and the nature of the setting in which the program is conducted (e.g. employment opportunities in the community). In correlational studies, the evaluator not only is interested in describing outcomes but adds another dimension by investigating relationships between outcomes and characteristics of students, the nature of students' educational experiences, or the nature of the environment or setting in which the educational program is conducted. Note, however, that evaluative studies that are intended to investigate relationships between outcomes and inputs do not necessarily result in conclusions denoting cause-effect relationships.

If an evaluator wishes to determine the effects (outcomes) produced by a particular educational program, we move to the third category. Here the evaluator is not satisfied with a description of outcomes only; nor is he satisifed in knowing what outcomes accompany certain program or environmental variables. What he is interested in are the outcomes produced by the educational program. If the evaluator is successful in establishing cause-effect relationships, he will also have described outcomes as well as established the degrees of relationship between outcomes and the type of program offered. Evaluative studies that claim to establish cause-effect relationships offer many challenges but have the potential to yield more valid findings than do studies designed primarily to describe or investigate relationships.

Comparison Groups. Another important but related concern which cannot be ignored in interpreting the findings of evaluative studies is illustrated by the question: With what group or groups

was "our group" compared? In many instances evaluative studies of vocational education programs do not compare the performance of persons who complete no vocational education program. In one-group studies of this nature we compare what are construe to be outcomes of the program with what we think the outcomes should be. Actually, this type of evaluative study is a case study that has as its primary intent the description of characteristics, capabilities, or performance of persons who have been exposed to the program. The evaluator who makes claims other than description for evaluation studies of this nature takes some rather dangerous risks.

A somewhat better alternative is to compare enrollees' characteristics, capabilities, or level of performance after completing an educational program with corresponding measures before they enrolled in the program. This is the familiar before-after study in which we compare enrollees after completing an educational program with themselves before or at the time of enrollment. Evaluators must be alert to factors that will be mentioned shortly which qualify, if not threaten, interpretations that can be placed on the findings of before-after studies.

The major concerns in interpreting evaluation findings whether or not a comparison group is used and the nature and characteristics of the group or groups with whom the outcomes of vocational education programs are compared. If we are to move beyond description with any degree of assurance that outcomes can be attributed to the educational program in which students participate, some provision must be made for the use of comparison groups, or control groups if you prefer the parlance of experimentation.

Alternative Explanations for Outcomes Observed

The main concern in evaluating educational programs is to determine what outcomes accompany or are produced by the programs. The problem is to identify some factors which have the potential for affecting program outcomes that, in turn, might be mistakenly interpreted as outcomes produced by the educational program. We are concerned with factors which threaten the truthfulness (validity) of evaluation findings. These possible threats offer plausible alternative explanations to the hoped-for conclusion that the educational program produced (caused) the outcomes observed.

How Students Are Selected. When outcomes of vocational education programs are compared with outcomes of alternative programs, care must be taken to avoid attributing differences in program outcomes to differences in programs when, in fact, differences in outcomes may be influenced more by differing characteristics of students than by differences in the nature of the educational programs. We know there are characteristics of people which are significantly related to performance regardless of whether an individual receives specific instruction or not. If students electing vocational education courses differ on some or all of these characteristics from students not electing vocational courses or from students electing alternative avenues to occupational preparation, evaluative studies comparing the capabilities or performance of persons completing vocational programs with persons completing other programs are misinterpreted when differences or lack of differences in outcomes are attributed solely to the nature of the educational program. Differences or lack of differences in program outcomes could be influenced, or perhaps accounted for, by differences between the groups of students on characteristics which are related to occupational performance.

Evaluative studies comparing the outcomes of vocational education programs with alternative programs that ignore or overlook the fact that enrollees in the various programs may also differ have high potential for yielding findings which may not be valid. The evaluator rarely has control over which students elect what program. Enrollees "self-select" themselves into the educational program. The same attributes, interests, and aspirations which lead students to select or not select a particular curriculum may also be the characteristics which enhance or impede outcomes independently--or in spite of--the curriculum in which they enroll. The "true experimenter" takes care of this problem through random assignment of students to programs which, within known statistical limits, achieves equality of the groups prior to enrollment. Evaluators of educational programs rarely have this option.

Contemporary History (Current Events). Another possible alternative explanation has to do with happenings in the students' environment, other than the educational program, which might contribute to favorable ratings on the criteria used to assess program effectiveness. Again there is the risk that findings of

the evaluation will be attributed to the educational program when, in fact, other events or experiences of students while enrolled in the program may have directly influenced outcomes. This possible threat to valid findings is labeled contemporary history, that is, events other than the educational program occurring from the beginning of the educational program to the time outcomes are assessed which may influence outcome criteria independently of the educational program. The threat of contemporary history is real to a valid interpretation of findings yielded by evaluative studies which do not involve comparison groups.

Disregarding Persons Who Do Not Complete the Program. Frequently, evaluative studies only involve persons who have successfully completed the educational program being assessed. If a considerable number of persons entering the program leave or drop out during the duration of the program, then it is obvious that an evaluation which involves only those who successfully complete the program may give a less than accurate (valid) picture of the outcomes of the program. Evaluative studies which only involve students who complete a program run the risk of misinterpretation if there has been considerable mortality of enrollees during the conduct of the program. So outcomes attributed to the program may, in effect, be due in large part to the fact that measures are only made on persons who are most capable as demonstrated by the fact that they completed the program.

Normal Growth and Development. I propose that any program designed to teach children ten to eighteen months of age to walk will be highly successful. This illustrates another possible threat to a valid interpretation of evaluation findings. It is obvious that most children between ten and eighteen months of age are going to learn to walk whether they participate in a formal training program or not. We must be alert to those situations where outcomes which we like to attribute to an effective educational program are not simply the result of normal maturation and growth of students. Maturation poses a threat that cannot be overlooked in long-term educational programs. Over a period of two or four years students are going to change considerably in physical, psychological, and emotional attributes which contribute to occupational success whether they are enrolled in an occupational education program or not.

Measurement and Observation. We know that the results of evaluative studies can be influenced by the type of measurements and observations used as well as by the persons who make the measurements and observations. In many respects it is difficult for persons responsible for the conduct of a program to evaluate the program objectively. There is always the temptation to use outcome measures which we think will indicate success; there is always the temptation, when ratings involve a great deal of judgment, to error in the direction that is favorable to the program. Care must be taken to design evaluative studies so that the outcomes claimed are not the result of biased and incomplete measurement and observation.

Test-wiseness. By taking tests we learn how to achieve higher scores on subsequent tests. When evaluative studies involve testing of students before and during the educational program, students may be learning how to achieve higher scores on tests at the end of the program because they become more test-wise. This tends to happen with attitude tests particularly and with achievement and performance tests when enrollees are not in an environment where testing is a common occurrence. Evaluative studies involving only one group of students where the same or similar tests are given before and after the educational program are particularly vulnerable to this possible alternative explanation for a finding that the program produced the outcomes described.

Extreme Cases. As a general rule, the more extreme we are on one measure today, the less extreme we will be tomorrow or sometime in the future on the same or a related measure. For example, students who score extremely high or extremely low on a test one day will tend to regress downward or upward on another test another day. One factor operating to produce that result is statistical regression. There are two situations where evaluators should be very sensitive to statistical regression as a possible threat to valid findings. When educational programs are provided students who score at the low end on criteria used in selecting participants, the students have no way to go but up. Evaluators must be careful not to credit this improvement in performance or achievement solely to the effectiveness of the program for it is likely that statistical regression has been a factor also. The other situation involves evaluative studies where vocational education students are matched with college preparatory

students on variables such as academic achievement or I. Q. scores. Statistical regression works in a very subtle fashion in these cases to produce, almost invariably, significant differences between the two groups on subsequent tests regardless of the treatment received by students in each group.

To What Extent Can the Findings Be Generalized?

Once the interpreter's curiosity has been satisfied about the validity of evaluation findings, the next concern has to do with the extent to which the findings can be applied (generalized) to other groups, situations, measures of outcomes, and times. The assumption here, of course, is that the findings are valid, that is, the findings describe outcomes which we are reasonably sure resulted at least in part from the educational program being assessed.

To What Groups (Populations Do the Findings Apply?

Characteristics of Students or Trainees. In generalizing findings about the relative effectiveness or ineffectiveness of certain educational programs, care must be taken not to assume that results produced when students with a particular set of characteristics are enrolled will apply to another set of students with different characteristics. Findings resulting from studies of persons with particular characteristics can be generalized only to populations of persons of which the persons studied are representative samples.

Nonresponse. A severe limitation to the generalizability of evaluation findings resulting from follow-up studies is non-response. This is particularly true when data are collected through mail questionnaires. In many cases a sizeable number of persons fail to respond. Unless there is evidence to the contrary, it can be assumed that respondents differ substantially from those who did not respond, especially on some variables which are crucial to the outcomes measured. The result is that the findings cannot even be generalized with any degree of confidence to all persons enrolling in or completing the educational program.

To What Settings, Situations, and Measures Do the Findings Apply?

Frequently, well designed evaluative studies are conducted in a particular school system. So the user of evaluation findings

must be alert to the temptation to assume that a successful program in one situation will be similarly successful in a completely different situation or setting.

Description of Program. The user or interpreter of evaluation findings is helped a great deal in applying results of evaluative studies if the evaluator has described in a thorough manner the nature of the program offered. This enables the user of the findings to make some rather clearcut decisions as to what other types of education programs the results might apply.

Hawthorne Effect. We know that people react differently when they are aware that they are participating in a special program or when their actions and activities are being monitored. It is not uncommon for better designed evaluative studies to accompany new programs and pilot programs. In these cases, it is very likely that those participating in the program, teachers as well as students, are aware that they are being evaluated. Evaluative results produced in these situations and settings may not be representative of what could be expected in similar settings with similar students and teachers who are not being monitored or are not aware that they are participating in a special evaluative effort.

Test Sensitization. Results yielded by evaluative studies involving many tests before and during the educational program may not be applicable to situations where a massive testing program is not used. Tests prior to and during an educational program sensitize students to the content of the educational program. Similar results might not be forthcoming in situations where students are not sensitized to what is to come.

Outcome Measures. Care must be taken not to assume that valid findings pertaining to one group of outcome measures will also hold true for another group of outcome measures. For example, findings regarding the economic benefits of vocational education cannot necessarily be generalized to findings that would result if noneconomic outcomes measures were used as evaluative criteria. Evaluators should design studies which include measures pertaining to all appropriate objectives of the educational program.

Recapitulation

I have listed seven possible alternative explanations for evaluative findings which state or imply that the outcomes do, in fact, result from the educational program being assessed. These factors, or threats to internal validity (Campbell and Stanley, 1966), can produce results which can be mistaken for outcomes produced by the educational program being evaluated. It behooves the evaluator as well as the user of evaluation findings to be aware of these possible threats to valid findings. Evaluative studies that do not involve comparison groups are particularly vulnerable.

In addition to making decisions about the internal validity of evaluation findings, the interpreter must also make some decisions about the extent to which the findings can be applied to other groups, situations, measures, and times. I have discussed some factors which should be considered in making decisions about the generalizability of evaluation findings. Researchers label the questions about generalizability as threats to external validity (Bracht and Glass, 1968).

We must not accept without question the findings of evaluative studies. We must not assume that just because we observe certain outcomes accompanying educational programs that the outcomes always and completely result from--or are caused by--the programs. We must be aware of the threats to valid interpretation and generalization of evaluation findings. If all we can do is describe the competence and performance of persons who have completed and left programs, let's admit it and make no further claims. If we can show a high degree of relationship between outcomes and program inputs, let's try harder to see if there really are functional relationships. If we claim that a particular program produces certain outcomes, let's demonstrate that we are aware of and have attempted to account for other factors that could produce or qualify the claims we make.

References

Bracht, Glenn H. and Glass, Gene V. "The External Validity of Experiments." American Educational Research Journal 5:437-474; November 1968.

Campbell, Donald T. and Stanley, Julian C. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally and Company, 1966.