COMMENTS ON PROFESSOR GLASER'S PAPER ENTITLED "EVALUATION OF INSTRUCTION AND CHANGING EDUCATIONAL MODELS"

Arthur A. Lumsdaine

University of Washington

From the Proceedings of the

SYMPOSIUM ON PROBLEMS IN THE EVALUATION OF INSTRUCTION

University of California, Los Angeles
December, 1967

M. C. Wittrock, Chairman

Sponsored by the Center for the
Study of Evaluation of Instructional Programs

The research and development reported herein was performed pursuant to a contract with the United States Department of Health, Education, and Welfare, Office of Education under the provisions of the Cooperative Research Program.

CSEIP Occasional Report No. 15, September 1968
University of California, Los Angeles

COMMENTS ON PROFESSOR GLASER'S PAPER ENTITLED "EVALUATION OF INSTRUCTION AND CHANGING EDUCATIONAL MODELS"

Arthur A. Lumsdaine

In relation to the controversy between Glaser and Stake, I want to say that I am on Glaser's side with respect to the importance of behavioral objectives. It is impossible for me to imagine how we are going to make progress in evaluation unless we do a better job of providing a rationale, a logical foundation for what we are measuring. I think that the most important contribution of programmed instruction to date has been less the improvement of instruction per se as a process of teaching than as an engineering effort emphasizing what is required to derive reasonable, measurable, useable instructional objectives from general statements.

I would enter one caveat here. In seeking what outcomes to assess in the process of evaluation, we can concentrate too fully on stated objectives and fail to include an assessment of the extent to which unexpected outcomes may eventuate. That is, the ultimate criterion of whether an educational program is a good program is what it accomplishes. It may have had some unintended bad effects which need to be ascertained, if possible, though they certainly were not among the objectives of the educational planner. A program also may have some unexpected good effects. Thus, if we confine our assessment of the outcomes of an educational program to its effects on its specified objectives, we may be overlooking some extremely important effects.

There are several points on which I would like to comment briefly: One problem of prime importance, in addition to improving the technology of evaluation (conceived as the means of ascertaining the outcomes of educational programs), is to try to create a better market for assessment data. As it now stands, there is not much demand for evidence about the effectiveness of specific programs. Educational products are still sold on the basis of unsubstantiated advertising. Those who have tried to change this situation have sometimes become quite discouraged by the realization of this. It seems rather futile to create data about an educational product and its effectiveness if there is little inclination on the part of those responsible for the purchase or selection of such products to look at the data. There is need for an educational job for the educational administrator or the curriculum supervisor, the person that makes the purchase decisions, to teach him about the usefulness of data and the demonstrable effectiveness of programs in making educational decisions. This important problem of long-range education is not necessarily going to be accomplished by those who are concerned with the technology of evaluation as such.

I also ought to mention my conviction related to the work of the American Psychological Association and the National Education Association joint committee (Lumsdaine, 1965), on which several of us here participated: that as part of a viable technology of assessment of program measurements we also need standards for the adequacy of such data. The reason for the standards derives from the following sequence of events. First, you have programs

with no evidence of output. Then you say, "Let's have evidence or data about the effectiveness of the programs." But then, in absence of standards, you get cheap, fallacious kinds of evidence, statistically reported in an impressive manner, perhaps, but technically unsound, i.e., with respect to methods of control. There emerges a sort of Gresham's Law of data, in which bad data, being easier to obtain and more impressive to report than good data, tend to drive out good data. So, some kinds of standards are needed, such as those which have come to be taken for granted and which will be observed in papers reported in, say, Journal of Experimental Psychology--but which are far from being safely assumed in the data being reported by evaluations of educational programs and materials.

I would also like to emphasize a point on which Marvin Alkin will probably comment further: namely, that if we are going to try to use cost-effectiveness criteria, we have to be able to measure output in cost translatable terms. Lack of this is the big hang-up, as I see it, in any cost-effectiveness program at the present time. It is very hard to say what the economic significance is of the difference between an achievement score of 128 and one of 212, even if these are translated into normative standard scores. The dependent variables which we characteristically use for measuring the outcomes of education are not easily translated into cost terms.

This may be, however, a fortunate accident, because I think that such measures as test scores are probably not what we really ought to use as measures of educational output, anyway. That remark

could be easily misunderstood. Let me see if I can clarify it.

Often we need to know that the most important competence is not that achieved through an educational procedure immediately after instruction. Rather, what we need to know (and this will become increasingly important as knowledge multiplies) is how well the effects of education enable a person to relearn something that he has forgotten, or to get quickly up to current operational proficiency from a background of prior training. There is just too much for everyone to know to expect that people will have a complete repertoire of competences on tap at all time.

What we need can be described in part as a problem of transfer. It is to begin to assess proficiency in terms not of what the person can do now or at the end of instruction or what he retains a week or a month or a semester later; but, rather, in terms of the amount of educational effort, what is required to bring him up to proficiency from where his education to date has left him (or to bring him back up to it after he has forgotten.

This implies something like a "savings" measure. Although there are many problems in developing and using such measures, they are attractive as measures of the effects of education which have promise of being translatable into cost terms. This is also true, of course, of the measure of instructional time needed to reach a criterion (as opposed to difference in scores after a fixed time of instruction).

Let me turn briefly to a different point. One thing that is very important to recognize clearly is that there are great

differences in the procedures needed for different purposes of evaluation.

The needs of program evaluation (particularly for such purposes as program improvement) are quite different from testing to evaluate the potentialities or achievement level of individuals. To take one example: when we are concerned with the evaluation of programs, we have quite a different sampling task than when we are dealing with the evaluation of individuals. For the evaluation of individuals, to oversimplify a little, we need a small sample of items about a large sample of people. We want to give each person a score but are willing to base each person's figure of merit on a relatively small sample of items. But in the problem of program evaluation, the opposite is true. We now want evidence from a relatively small--adequate but relatively small--number of individuals on all relevant items. This is because we are assessing the program for the purpose of product improvement; so, it is very important for us to have this fine-grain differential knowledge of the successes and the failures of the program on each point it covers to detect its very specific failures and successes.

I have tried hard to think of something to disagree with Bob Glaser about. One point of partial dissent concerns the alleged lack of relation between intelligence and program effects. There are, in fact, numerous instances in which successful attempts have been made to relate measures of ability to differences in programs. I can think of examples because I have been involved in collecting, analyzing, and reporting such data, in studies in

which effects of instructional devices, such as films, were analyzed with concern for differential effects on individuals of greater and lesser ability, as measured by standard tests of intelligence. Great differences, in fact, were found between the effectiveness of programs as a function of such ability-test scores and educational level of adults (Hovland, Lumsdaine, and Sheffield, 1949).

These differential effects as a function of stratifying variables, such as educational level, IQ, or other measures of ability, are of the form of interactions talked about as an important outcome of our educational tests. However, there are two kinds of interactions. One kind is as follows: The effects of Program A are greater for Group X than they are for Group Y. Here the difference in the relative effectiveness of two programs, A and B (e.g., color versus black and white films, overt versus implicit response procedures, etc.) is greater for a segment of one group of the population, X, than it is for some other segment, Y. There are many instances where this is the case--where a particular instructional variable demonstrably makes more or less difference for, let's say, brighter students than for less bright students (cf. Hovland, Lumsdaine, and Sheffield, 1949, chapters 8 and 9).

However, the argument for the individualization of instruction rests in part on the assumption that there is a more powerful kind of interaction at work than the kind just described. This would be where not only is the difference between A and B greater for Group X than for Group Y, but A is superior for Group

X while B is superior for Group Y. This is a "reversal" kind of interaction that demonstrates, as a function of some population characteristics, that one program is better for certain persons, whereas other programs are better for other persons.

Now, if you search the literature, both on formal instruction and on attitudes, you will find very few such instances of reversible interactions documented by solid evidence. There are a few of them, and modesty forbids mention of the first that come to mind. But it is interesting that with all our talk about the importance of tailoring instruction to individual characteristics, we find so few instances of differential effects of this kind; so we can conclude that the program is tailored for a particular, not just individual, subgroup of individuals and is differentially effective for them.

REFERENCES

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. Experiments on mass communication. Princeton, N. J.: Princeton University Press, 1949.

Lumsdaine, A. A. Assessing the effectiveness of instructional programs. In R. Glaser (Ed.), Teaching machines and programmed learning, II: Data and directions. Washington, D. C.: National Education Association.