COMMENTS ON PROFESSOR TROW'S PAPER ENTITLED "METHODOLOGICAL PROBLEMS IN THE EVALUATION OF INNOVATION"

Eugene Litwak

University of Michigan

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

This 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.

CSE Report No. 32, May 1969
University of California, Los Angeles

COMMENTS ON PROFESSOR TROW'S PAPER ENTITLED "METHODOLOGICAL PROBLEMS IN THE EVALUATION OF INNOVATION"

Eugene Litwak

I should like to make some comments on specific points raised by Professor Trow and then advance a multi-model theory of evaluations with ensuing predictions as to what type of evaluation strategies might be ideal for twenty-four "generic" situations.

Objective and "Intuitive" Evaluation Techniques. Professor Trow provided a good case in point for Robert Stake's view (stated in his discussion with Glaser) that the degree to which we can provide good measures is not necessarily related to the importance of the objects we are trying to evaluate. As a consequence, when we do not have "objective" measures we may have to utilize crude evaluation techniques. Insisting on more objective measures may mean no evaluation at all or one which is quantifiable but a poorer predictor than quantitative judgments. Thus, Professor Trow points out that it is difficult to operationalize some of the goals of higher education--the notions of good citizenship and liberal education. These are goals which might be achieved 10 to 15 years after a person leaves college and which involve properties which are difficult to measure. The history of evaluation has been one where we have tried to introduce quantification into new areas. On the whole, this has

been beneficial. However, as this movement has gained success the dangers mentioned by Stake, and suggested in specific detail by Trow increase. Current evaluation specialists must increasingly ask themselves when to use quantitative techniques for evaluation and when to use more qualiltative techniques, rather than assume that invariably quantitative techniques are better. This argument must be differentiated from the one in the past where a hard core group resisted all systematic qualitative evaluation and another insisted on it. In the second part of the paper we will suggest specific evaluation procedures where people can make only a gross estimate of their goals.

Daily Effectiveness and Program Results--Two Types of Evaluation. Another point made by Trow illustrates something discussed by Lortie and Gage in their exchange. Trow points out that it is difficult for people to accept evaluations, especially when their jobs are at stake (e.g., when someone in a position superior to theirs is involved). A host of literature (including an article by Lortie) supports this view. In effect, what Trow suggests is that perhaps we should find a way to put evaluation in the hands of the people who are doing the job or in the hands of their colleagues. I think this touches upon the discussion between Gage and Lortie as to where evaluation efforts should be made.

I would suggest that there are two legitimate notions of evaluation that should be accepted. One is the notion of daily job

effectiveness. The individual uses the daily information to change his behavior and perform his job more effectively. As mentioned above, there does seem to be some evidence that such kinds of evaluations require a trusted colleague or individuals themselves to do the evaluation. The major problem with this kind of an evaluation is that the evaluator becomes too identified with the individual being evaluated and in situations of ambiguity is likely to orient the evaluation in terms of personal welfare rather than around the goals of the organization. This is commonly recognized in the cry from the "objective" outside evaluator. It is my view that this second type of evaluation is also necessary; I would call it an overall program evaluation. It does involve outside or "impartial" evaluators and the total administrative hierarchy. It is also characterized by the fact that it is not a daily evaluation but yearly or less frequent. This evaluation has all the problems raised by Trow--that people will find it difficult to accept--as well as the virtues of being able to take a hard look at what is being accomplished. It seems to me that we have two important problems with regard to evaluation and each of them requires a different kind of evaluation. I think that Trow has emphasized only one side of the issue. I would suggest that the evaluator must, in any given situation, make a diagnosis of the problem. Is he trying to find methods for getting teachers to improve their daily efforts through some systematic feed-back device, or is he interested in

overall program evaluation? Both are legitimate goals and at any given stage in an educational institution's development he may want one or the other or both stressed.

Hawthorne Effect and Social Engineering. My next point concerns what Trow referred to as the "Hawthorne" effect. His point is very similar (and paradoxically different) to the remarks made by Gage in his description of Stephen's work. Gage points out that most evaluations show that different school programs make little difference on the students' progress. By contrast, Trow points out that most experiments in education seem to work. These are not necessarily contradictory propositions since one is talking about experiments and the other about established school programs. What is similar about both of these propositions is that both Stephens and Trow suggest that the crucial underlying variable is teacher ability and enthusiasm. These are far more important than program variations. Within a school system teachers with outstanding abilities are randomly distributed among the programs. That presumably is why difference between programs means so little. Among the experimentors and the non-experimentors they are not randomly distributed. The experimentors are usually highly enthusiastic and able. That is why all experiments work.

I would agree that the Hawthorne effect is an important one and that we should concentrate on ways for maintaining it continuously. However, I think that there are also legitimate problems of social engineering which might explain why successful experiments cannot be

translated into successful school programs. It seems to me that often experiments have many hidden complexities aside from ability and enthusiasm of the investigator which the investigator cannot translate to a system-wide basis; sometimes because the investigator is not aware of them, often because there is a lack of knowledge as to how to introduce innovation into a system (both the letter and the spirit of the innovation) and often because the system cannot put the kind of resources used in the experiment into the general application and what emerges is a watered down version of the experiment.

I would be somewhat pessimistic about our educational establishment if indeed all that was involved was a "Hawthorne" effect, because I doubt very much that mass institutions can find sufficient people of the high calibre and degree of enthusiasm suggested by such analysis. I would therefore suggest that we concentrate in addition on the organizational basis for accepting innovation of all kinds rather than how to maintain involvement at the highest pitch.

Towards a General Theory of Evaluation. I think throughout this conference there has been a questioning as to whether there is one ideal form of evaluation which holds in all situations or whether we have different strategies of evaluation for different situations. Glaser raised this point quite clearly. I would opt for the latter point of view and would now like to review some of the elements which would have to be considered and the differential evaluation techniques they

imply. The variables I am suggesting as being generic and their relationships to evaluation strategies are as yet very primitive. However, I do want to go beyond the platitudinous statement that different situations require different evaluation techniques. With this limitation in mind, the following are some of the factors which can be used to differentiate all situations and as a consequence suggest differential evaluation techniques.

Current State of Knowledge. The "classic" evaluation technique is very close to the pure experiment or "classic" planning strategies. The suggestions in all cases tend to be the same. First specify the goal then the alternative strategies (i.e., teaching procedures) for reaching this goal. All the evaluator has to do is to measure the children before the new program is introduced, measure him after and decide which if any of the programs show the most marked difference. Assumed in this analysis is the ability to define one's goals clearly (measure their achievement) as well as to specify the range of alternative means. Professor Trow has pointed out that it is often difficult if not impossible to measure one's goals or even to specify them clearly. He might have also added that it is often difficult if not impossible to specify alternative means. There are various reasons for this, (e.g., there is not enough time, it costs to much, etc.). However, in this section I want to stress one reason--the state of knowledge. Is there any theory which systematically suggests what are the best evaluation strategies when we have incomplete

knowledge? Most of them start out with the premise that before evaluation can begin we must have excellent states of knowledge. The work of Dahl and Lindbloom and more recently that of Lindbloom on decision making strategies provide some useful alternatives. They suggest in situations where things are going reasonably well in the sense that there are no major calamities, that one use an incremental strategy. This implies introducing innovations which tend to be simply monotonic projections of past historical trends and which are reversible. This often means small innovations. If nothing major happens then one continues this process. Still assuming that one has only a gross specification of goals and little knowledge of alternative means, they suggest that an alternative strategy be used when the situation is bad, (as judged by gross qualitative evaluation). Thus, a major depression or the clear sense of the community that the school procedures are not working well in the inner city would be cases in point. In this situation they suggest a "calculated risk" strategy. The main point of this strategy is that one is to depart as radically as possible from past historical trends and pay less attention to the reversibility of the innovation. The reasoning behind this directive is that where things are going very badly, little can be lost and much gained by radical shifts in methods.

The important point to be stressed is that they are suggesting "rational" strategies in situations where we have incomplete knowledge.

If their arguments are correct they also suggest criteria for evaluation under incomplete states of knowledge. What they are saying is that the evaluator need make only the grossest qualitative assessments about goals in situations where goals cannot be clearly specified because of lack of knowledge. Thus, the college faculty must make a decision right now as to what constitutes requirements for a liberal arts degree. Yet the goals they seek to achieve (such as good citizenship and the humanitarian man) cannot be measured right now with any degree of accuracy. At this point, Dahl and Lindbloom would be suggesting that the evaluator only has to make, in conjunction with his client, a qualitative judgment as to whether liberal arts programs have failed or not. If he feels that they have not obviously failed in the sense that there is no general complaint or he has some positive general assessment, then he might adopt the incremental approach. This means he should measure any innovation on three criteria--does is fit within the historical trend, is it reversible, does it have any consequence based on the same kind of generalized judgment which can be thought of as definite failure or success? Alternatively, if the initial assessment is that the current situation is very bad then the evaluator uses the "calculated risk" as the basis for setting evaluation criteria. In both cases where historical data are not available the evaluator might utilize as his comparison group other institutions engaged in similar work and in similar circumstances.

To summarize, what is being said is that where one has relatively complete information as to goals and means then one can use the traditional "experimental" before and after evaluation approach. However, where one lacks knowledge, then one uses only gross judgments on goals and turns one's attention to the evaluation of a given approach as being historical on or off the trend line as well as judging the reversibility of the innovation. The more completely one can develop a theory of decision making under circumstances of differential states of knowledge, the more confident one can be about having a general theory of evaluation that fits the problems that often confront evaluators (e.g., how to evaluate with incomplete knowledge).

Economic Manpower Scope of Evaluation. Another problem which emerges in evaluation is the scope of the evaluation procedures. Should we jump into an evaluation of total systems or should we first evaluate small experimental programs? It seems to me that one might move towards small experimental laboratory evaluation procedures where one has good knowledge (operational measures) of goals and alternative means but little knowledge as to their relationship. A small laboratory based evaluation situation permits the investigator to engage in all kinds of variations with minimal concern for costs. Thus, the general rule would be that where one is suggesting the use of very costly evaluation processes and where one has high states of knowledge on means and goals but not their relationship to each other, the evaluator moves toward a small experimental model. By contrast, where he has low cost

processes and either high or low states of knowledge he might want to utilize large scope evaluation procedures (e.g., large field experiments or surveys). This discussion bears directly on the point that Alkin was making. Where a technique was extremely costly the evaluator might either restrict it to small experimental situations or even say it is not worthwhile studying even if it were the most successful. Thus, a teaching method which says that there must be one teacher for every child in the school might be the most successful teaching technique, yet one which we would not bother to evaluate or evaluate in a laboratory-like situation since even with the optimal effectiveness, the costs would be too high for any system to undertake.

Controllability of Independent Variables, and Experimental Versus Survey Procedures. Another factor which obviously affects the evaluation procedure is the controllability of the independent variable. Often in the field of education as well as in social sciences in general it is difficult to control our independent variables. We are often in the position of astronomers rather than laboratory experimental physics. For instance, we are often in the position of looking at two schools, one which has a close school-community relationship and the other which does not. We want to see what difference this makes for the child's reading skills. However, we are not in a position to get the schools to alter their procedures systematically. If we are fortunate and can spot these incipient experiments before hand, we can do some panel

he must deal with, and the generalized community support for such a program.

Any time the stimulus is a complex one (i.e., consisting of many independent variables with some causal links to each other as well as to the dependent variable) the kind of model that Gagné was suggesting would be difficult to undertake. It would involve an intolerable number of controlled experiments and might yet miss the overall causal links between independent variables. In such situations one might well move to a very large survey or panel study which permitted, in a relatively short period of time, an examination of many different combinations of variables. This might not have the logical eloquence that is suggested by the pure experiment but it has the virtue of providing useful information in a reasonable time.

There is nothing said so far which is very new. However, I would suggest that two things derive from the above analysis which might be viewed as more controversial. First, on the basis of the reasoning I have just gone through, we should forego the notion that there is one ideal mode of evaluation and move towards the concept of a multiple model. In fact, this is what most evaluators are now doing, and what I am suggesting is that rather than viewing this as a departure from an ideal norm we view it as an ideal state. This in turn leads to the second point; is there some theory which states what type of evaluation processes are ideal for the various situations which confront

evaluation. Can we show that there are really a limited number of dimensions which characterize most situations we have to evaluate? If so, we have a finite number of models of evaluation procedures rather than an infinite number. The specification of the basic dimensions for classifying situations as well as their evaluation outcome would constitute a multiple model theory of evaluation. What I have done in the above section of this paper is suggest some of the obvious starting points for such a classificatory scheme as well as some of the evaluation outcomes. To make this point quite clear, these dimensions must now be simultaneously considered and the forms of evaluation which ideally emerge from this simultaneous interaction specified.

Table one presents in tabular form my first approximates of a multiple model theory of evaluation. This theory is based on all possible combinations of the following simple principles.

I. Complete knowledge of ends and means permits true experimental evaluations and the purposeful sampling of individuals where necessary, (e.g., a priori matching groups).

   Incomplete knowledge of ends and means generally precludes the use of experimental designs--requiring either survey or panel analysis type instruments and requiring random selections of populations. Where the overall lack of knowledge is coupled with the gross evaluation that the situation is alright then

Table I

RELATIONSHIP OF SOCIAL FACTORS TO TYPE OF EVALUATION

Complexity of the Stimulus	Control of the Stimulus	Knowledge of Ends and Means Very Good (Good Operational Measures)		Knowledge of Ends and Means Very Poor (Poor Operational Measures of Ends)	
		Costly to Collect Data	Not Costly to Collect Data	Costly to Collect Data	Not Costly to Collect Data
Simple Stimulus (one or two independent variables not causally related)	Complete Control	1. Few small laboratory experiments	7. Small field experiment	13. Logically not possible to have no knowledge and complete control	19. Logically not possible to have no knowledge and complete control
	Partial Control	2. Small panel study with highly stratified sample	8. Medium sized panel study with stratified sample	14. Small, simulated panel, random sample, trend and reversibility analysis	20. Medium simulated panel, random sample, trend analysis reversibility analysis
	No Control	3. Small survey with highly stratified sample	9. Medium sized survey with stratified sample, e.g., around natural experiment	15. Small survey, random sample, reversibility and trend analysis	21. Medium survey, random sample, reversibility and trend analysis
Complex Stimulus (many independent variables related in a causal sequence)	Complete Control	4. Many laboratory experiments coupled with survey data (e.g., questionnaires to respondents)	10. Large field experiment coupled with survey analysis of respondents	16. Logically not possible to have no knowledge and complete control	22. Logically not possible to have no knowledge and complete control
	Partial Control	5. Medium panel study with stratified sample	11. Large panel study with stratified sample	17. Simulated panel (medium), random sample, trend analysis, reversibility analysis	23. Large simulated panel, random sample, trend analysis, reversibility analysis
	No Control	6. Medium surveys with stratified sample	12. Large survey with stratified sample e.g., natural experiments	18. Medium survey, random sample, trend analysis reversibility analysis	24. Large survey, random sample, trend analysis, reversibility analysis

the evaluator seeks comparative data--i.e., either historical or not--with the goal in mind of judging any innovation in terms of its continuity and reversibility. Where this lack of knowledge is coupled with a gross evaluation that the current state is very bad then comparative data is examined to see how far the new innovation departs from the old.

II. Where the evaluation process is costly (in terms of time, manpower, or general economic resources) then small laboratory evaluation procedures are desirable. Where the evaluation process is not costly then large scale surveys or field experiments are possible.

III. Where the evaluator has complete control over the stimulus he can use experimental designs, where he has only partial control he needs to use partial experimental designs like panel analysis, while where he has no control he must use techniques like survey analysis.

IV. Where the stimulus to be examined is very simple it provides an ideal situation for small group experiments, whereas if the stimulus is very complex (there are many independent variables and they are related to each other in a causal sequence) then large surveys or panel studies will generallly be necessary.

With this in mind, we can look at cell number 1 in our table. According to our multiple model theory of evaluation this is the situation

where the evaluator should use a small experimental laboratory study to do his evaluation because he has fairly good knowledge of the means and ends, the stimuli (means) are very simple, it would be costly to do the evaluation on a large scale, and he is able to control the stimulus. By contrast, if the evaluator is in a situation described by cell 24 he would use large scale surveys with random samples. This is true because he lacks knowledge to operationalize the ends, he cannot control the stimulus, he assumes the stimulus is complex, and he can collect much data inexpensively. These conditions prevent him from setting up an experimental laboratory evaluation or even seeking a natural experiment. At the same time they put a premium on gathering much information (e.g., complex stimulus, lack of knowledge, and low costs).

The reader will note that cells 13, 16, 19, and 22 are all considered to be logically impossible. It is argued that in situations where there is incomplete knowledge of ends and means, one cannot (by definition) control the means (stimulus). If we examine cell 12 we find an interesting mixture which in turn suggests a slightly different kind of evaluation method. This is a situation where there is knowledge of ends and means but where the investigator cannot control the stimulus. This is a typical problem of astronomy. In addition, the stimulus is very complex which tends to suggest the use of large survey and this is further reinforced by the low cost of the evaluation. However, this survey can differ from the survey discussed in cell 24

because here the investigator has much more knowledge of the ends and means. He can put this knowledge to use by his sampling procedures. He can either sample to insure that he has incorporated natural experiments or he can stratify his sample to insure that he has relatively equal numbers of cases for all of his major variables. Cell 6 is like cell 12 except we now have a situation where the costs of the evaluation are high. In such circumstances the size of the sample will probably shrink so we now have a medium rather than a large survey. Cell 11 is also like cell 12 but it differs in that the investigator has some control but not complete control over his environment. This suggests that he might be on the scene before a natural experiment is begun and thus he might be able to get before and after measures and do a panel analysis though not have a true experiment. Cell 5 is just like cell 11 but involves a more costly evaluation so we would suggest the chief thing differentiating them would be the size of the panel study. Cell 10 is like cell 12 but here the investigator has control over his environment. This permits an experiment but the large number of variables would suggest that he might not be able to do all possible experiments nor would many single experiments necessarily unravel the interactions between the independent variables. Since this cell also states that we are not dealing with a low cost situation, it would seem to us that a large field experiment coupled with much interview data would be appropriate. The experimental design will permit one to test out some of the variables through experimentation while the use of the

survey and consequent panel analysis would permit one to use statistical analysis to deal with the more obscure variables and the more intricate set of interactions. Cell 4 would be like cell 10 but for the increased cost. This may mean smaller field experiments or the use of many laboratory experiments as the chief evaluation procedure. The reader will recall that we said that cell 1 was the ideal situation for a small laboratory experiment. We think that cell 7 would be the ideal situation for a small field experiment. It is exactly like cell 1 but there are little costs in doing the field experiments so it should be done because it often means one less inference for the evaluator (e.g., will the laboratory results hold in the field). The reason that this field experiment can be small whereas cell 10, which is very close to cell 7, must involve large field experiments, is because cell 7 has a single or simple stimulus. The reasoning for cells 2 and 8 follow those for 5 and 11 with the difference being in a simple rather than complex stimulus. Similarly, 3 and 9 follow 6 and 12.

If we now examine the opposite side of the table where we have incomplete knowledge, it has already been noted that cell 24 differs from cell 12 (which matches it except there is complete knowledge) in having a random sample rather than a purposeful sample. In addition, this theory suggests that where incomplete knowledge is coupled with a positive gross evaluation of current activities the

evaluator will utilize his statistical techniques to look at the method being evaluated historically (through retrospective questions) or comparatively with similar organizations and all assessments will be guided in terms of their "fit" to historical or comparative trends. In addition the methods will be evaluated in terms of their reversibility. In contrast, if the gross evaluation is that the current situation is very bad, then the evaluator, using the same comparative data, will see how far the innovation departs from the historical or comparative standards. Cell 23 is almost the same as cell 24 but here there is some partial state of control over the stimulus. However, it suggests that the control may not be quite as great as cell 11 which is the same except for the knowledge base. Therefore, it is suggested that here we might have some kind of simulated panel study--through use of cohort analysis and possibly two cross sectional surveys taken at two different periods of time but not with the same people. Using statistical devises one can have a simulated panel design. Cell 18 would be like cell 24 but requires a medium sized survey because of the cost factor and cell 17 would be the same as cell 23 but smaller in size because of the cost factor. Cell 21 would like cell 24 but because of the assumed simple stimulus would require a smaller sample size while cell 20 would be a smaller version of cell 23 for the same reasons. Cell 14 would, because of cost, probably be like cell 20 but even smaller while cell 15 would be like 21 but smaller.

This now completes the provisional analysis. We have generated almost 24 different types of evaluation techniques. No attempt is made to argue that this is where an evaluation theory will eventually lead. However, it does illustrate in more detailed terms what we mean when we say there must be a multi-model theory of evaluation. Hopefully, this initial formulation, crude as it may be, will encourage others to pursue this inquiry more deeply.

OE 6000 (REV. 9-66)

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
OFFICE OF EDUCATION

ERIC REPORT RESUME

(TOP)

ERIC ACCESSION NO.

CLEARINGHOUSE ACCESSION NUMBER

RESUME DATE 5-1-69 | P.A. | T.A.

IS DOCUMENT COPYRIGHTED? YES ☐ NO ☒

ERIC REPRODUCTION RELEASE? YES ☒ NO ☐

001

TITLE

100 Comments on Professor Trow's Paper Entitled "Methodological Problems
101 in the Evaluation of Innovation."
102
103

PERSONAL AUTHOR(S)

200 Eugene Litwak

INSTITUTION (SOURCE) | SOURCE CODE

300 UCLA - CSE

310 REPORT/SERIES NO. CSE Report No. 32

OTHER SOURCE | SOURCE CODE

320 Symposium on Problems in the Evaluation of Instruction, Dec. 1967

330 OTHER REPORT NO.

OTHER SOURCE | SOURCE CODE

340

350 OTHER REPORT NO.

400 PUB'L. DATE 5-69- | CONTRACT/GRANT NUMBER OEC 4-6-061646-1909

PAGINATION, ETC.

500
501

RETRIEVAL TERMS

600
601
602
603
604
605
606

IDENTIFIERS

607

ABSTRACT

800 Litwak proposes two distinct notions of evaluation: a facilitative evaluation
801 which at aims at the improvement of everyday operations, and a more
802 formal evaluation which seeks to evaluate the entire program. Concerning
803 the use of surveys in evaluation, his paper stresses the need to classify
804 events in terms of their complexities and then to designate appropriate
805 methodological procedures.
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822