DOCUMENT RESUME

ED 073 138 TM 002 382

AUTHOR Cook, Thomas D.
TITLE Secondary Evaluations.
PUB DATE 8 Sep 72
NOTE 13p.; Paper presented at the American Psychological Association meeting, Hawaii, September 8, 1972

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Comparative Analysis; *Data Analysis; *Evaluation Methods; *Program Evaluation; Technical Reports

ABSTRACT

Secondary evaluations, in which an investigator takes a body of evaluation data collected by a primary evaluation researcher and examines the data to see if the original conclusions about the program correspond with his own, are discussed. The different kinds of secondary evaluations and the advantages and disadvantages of each are pointed out, and the particular advantage of one kind of secondary evaluation which is likely to be relatively comprehensive in scope is discussed. What secondary evaluations can and cannot accomplish is noted, and some specific steps that could be taken to increase the frequency and quality of secondary evaluations in the future are outlined. (Author/DB)

Do be revised

Paper presented at APA
Hawaii, September 8, 1972
Comments appreciated

SECONDARY EVALUATIONS

Thomas D. Cook

Northwestern University

Introduction.

My topic today concerns the scope of evaluation research, and I shall approach the topic by way of secondary evaluations. We are most of us familiar with primary evaluations in which a contract is issued to a research organization which then evaluates some program and issues a report about the program's effects on its recipients. Some of you may be less familiar with secondary evaluations in which an investigator takes a body of evaluation data collected by a primary evaluation researcher and examines the data to see if the original conclusions about the program correspond with his own. What I want to discuss right now is; first, the different kinds of secondary evaluations and the advantages and disadvantages of each; second, the particular advantage of one kind of secondary evaluation which is likely to be relatively comprehensive in scope; third, I want to place secondary evaluations into a realistic perspective as to what they can and cannot accomplish; and finally, I will outline some specific steps that could be taken to increase the frequency and quality of secondary evaluations in the future.

Types of Secondary Evaluation.

Perhaps the most frequent form of secondary evaluation is the critical examination of a research report. This takes place when a report is examined, criticized, and some tables are reworked in order to see if the report's conclusions stand after reanalysis. There have been many examples of this, the most famous being the critique of the "Authoritarian Personality" by Hyman and Sheatsley,[1] and we have probably all experienced the same kind of evaluation in the anonymous critiques of articles that we have submitted to journals. This form of secondary evaluation is of distinct but limited utility. It is useful because it serves as some measure of quality control, but is limited because it does not involve the direct reanalysis of the data.

The next most frequent form of secondary evaluation could be called the data-bound reevaluation of a research report. This involves obtaining a body of data and then reanalyzing it to ascertain whether the original conclusions can be corroborated. Examples of such work include the reevaluation by Elashoff and Snow[2] of the book "Pygmalion in the Classroom" by Rosenthal and Jacobsen,[3] and the chapters in Mosteller and Moynihan's book[4] on the Coleman Report.[5] The basic aim of such secondary evaluations is to ask: Do the Rosenthal or Coleman data support the conclusions drawn by the respective investigators? This is obviously an important question if the original conclusions have important implications, and having the data on hand permits investigators to explore the data in a way that is often foreclosed to the primary investigator who has his deadlines to meet. However, there are two salient problems with this kind of secondary evaluation. One is that the analyst conceives of his work in terms of the conclusions of earlier reports and he may only explore questions

relevant to these conclusions. Thus, he may be less likely to ask whether the conclusions reflect the most useful questions to be asked of a program and he is less likely to ask whether these new questions can be answered either with the data on hand or with relevant reports by others. Let us consider the evaluations of Sesame Street by Ball and Bogatz[6] in this light. In setting out to evaluate the two reports by the ETS team, one would tend to ask their primary research question: "If children view Sesame Street, do they learn from it?" One would be less inclined to ask "What is Sesame Street's national impact?" for this question is not explicitly dealt with in the ETS reports and could only be answered by considering the ETS data as well as survey data about the show's national viewing audience. The second major problem with evaluating the conclusions in a research report is the degree of conflict between primary and secondary analysts that this strategy could engender. After all, the secondary analyst aims to see whether the primary analyst has done a good job, and it is almost inevitable for reasons given below that the secondary analyst will find holes in the previous work. If secondary analysts are dependent on the goodwill of primary analysts for the release of data, we might expect this goodwill to decrease over the years if we continue to evaluate evaluations. To counteract this, we shall either have to institutionalize other forms of access to data for secondary analysis purposes or we shall have to come up with a better model of secondary evaluation or we shall have to do both.

Perhaps a better, but still imperfect, model of secondary evaluation is the reevaluation of a program. Here the stress is on using the data to design an evaluation from scratch without regard to how the primary analyst originally analyzed his data. Thus, one's aim would be to evaluate Sesame Street anew, instead of to evaluate the ETS evaluations, and one

would present one's conclusions about Sesame Street with as little reference as possible to the original report. The problem here is that it is impossible to ignore the original report especially where the primary and secondary analyses produce conflicting results. Then, it is encumbent on the secondary analyst to reexamine the primary analysis and, where he thinks the latter is incorrect, he has to point this out, thereby evaluating the original evaluation. Hence, the advantages of evaluating programs instead of evaluations is a relative and not absolute advantage, and the potential for antagonism that is built into the role relationship of most primary and secondary analysts will be reduced but not eliminated. Obviously, any antagonism can detract from the quality of the secondary analysis because the secondary investigator is frequently dependent on the primary investigator for important information which is not contained in research reports in the detail that is desired.

The major advantage of program reevaluations (as opposed to evaluation evaluations) is that they force the secondary analyst to begin the evaluation process from the very start. Most primary or secondary evaluations that I have read are relatively sparse in the following sense. All deal with the issue of whether the program being evaluated causes statistically significant changes in its recipients; some deal with the issue of whether the magnitude of effects differs according to characteristics of program members; some deal with the unique contribution of separate parts of the typically global treatments; some deal with whether the program causes gains that are socially significant; other evaluations deal with some of the unintended negative and positive side effects of the program; some deal with the number and social characteristics of persons reached by a program; some deal with the dollar cost of the program per program member per year;

some deal with the differences in dollar costs for different kinds of program recipients; and some deal with the social values that may be advanced or hindered by a specific program. But no evaluation puts all these elements together to give a comprehensive picture of a program's effects on its members and on society at large. The more comprehensive picture can radically alter one's global evaluation of a program. For example, if one finds that an autotelic teaching machine teaches some reading skills but costs three times more per pupil per hour, this influences one's evaluation of the machine; or if one finds that the drilling of learning causes IQ gains but a decreased desire to graduate from high school, this should influence one's evaluation of drilling; or if one finds that the highest pretest scorers on, say, an aptitude test, learn more from a compensatory education program, this should force us to consider whether it is more important to "open doors" for a few or to "raise the floor" of all the disadvantaged. I feel that secondary evaluations of a program make evaluations relatively more comprehensive because the analyst is freer to ask whatever questions he wants about a program, whether these reflect the program's objectives, or claims made for a program, or the possible effects of the program on some national problem; also, the secondary analyst has more time for his task than the primary analyst; moreover, he does not have a restricted commission only to evaluate whether program recipients experience some statistically significant benefits from a program; and finally, he is freer to examine other bodies of relevant data in reaching conclusions, especially data about issues that were not brought up in the primary evaluation.

Making Evaluations More Comprehensive.

Let me illustrate the advantages of the secondary analyst who sets

out to evaluate a program from scratch by outlining the question-asking process that we went through in evaluating Sesame Street. The first problem was to decide the basic questions that were to guide the evaluation. We knew that the show's objectives would give us clues, though we also knew that they would not define the universe of intended or unintended effects of the program. Ms. Cooney stated in 1968 that the show was intended to stimulate "the intellectual and cultural growth of all preschoolers, particularly disadvantaged preschoolers."[7] We realized from this that we would need to explicate and search for indices of different kinds of intellectual and cultural growth and we also realized that we would have to test whether preschoolers of different social groups learned from Sesame Street if they viewed the show. We were less sure of what the phrase "particularly disadvantaged preschoolers" meant. Did it mean that a special effort would be made to reach the disadvantaged? Did it mean that Sesame Street would be vindicated if disadvantaged viewers learned irrespective of how much advantaged children learned? Or did it mean that the achievement gap between the advantaged and disadvantaged would have to be narrowed? In the end, we decided to test both whether the disadvantaged learned from Sesame Street and whether the knowledge gap was narrowed. We decided on the latter question, first, because in her original request for funds, Ms. Cooney discussed Sesame Street's special focus on the disadvantaged in the context of the achievement gap; second, because Ball and Bogatz had claimed that Sesame Street was narrowing the gap; and third, because the gap is considered a grave national problem and any widescale educational innovation with a special emphasis on the disadvantaged will be relevant to this national problem. Once we had decided to focus on whether all groups learned and whether the achievement gap was narrowed, we had to ask how

these two objectives could be met. One way would be if all groups learned equally from equal amounts of viewing and the disadvantaged were more likely to view; and another would be if the disadvantaged learned more from equal amounts of viewing and if they viewed as much as or more than the advantaged. Thus, our basic questions were set and they implied a different evaluation than that of Ball and Bogatz. We were relatively less interested in the question "If children view Sesame Street, do they learn?" and we were relatively more interested in the question "What is the impact of Sesame Street on the national problems of raising the academic and cultural knowledge levels of preschoolers of all kinds and on the national problem of narrowing the academic achievement gap between the advantaged and disadvantaged?"

It is not clear whether Ball and Bogatz would have been allowed to pose the evaluation questions in the way we did even if they had wanted to. Their work was commissioned by the Children's Television Workshop which in its prebroadcast years did not expect to be the great national success it was. Sesame Street was initially presented to the public as a modest experiment, and officials at CTW probably thought it unrealistically grandiose to think of the program as ameliorating long-standing national problems. Moreover, the national perspective requires information about the size and social composition of the show's audience. CTW officials were initially sensitive to the fact that the show was broadcast on educational television channels where the better educated were overrepresented among regular viewers. Thus, they believed at the outset that the chances of the disadvantaged viewing as much as the advantaged were slight and that the chances of them viewing more than the advantaged were even slighter. A secondary analyst is not dependent upon program officials as a primary analyst is,

and he is freer to decide on his own version of the major evaluation questions.

The next stage in an evaluation, whether primary or secondary, is to test whether a program causes statistically significant effects on the variables of major concern. This is a major thrust of most primary evaluations, but even here the secondary analyst has advantages. For one, he is made aware of any unique new designs or approaches to data that the primary analyst has developed to take account of special problems; and for another, he is made aware of some problems that exist and for which no adequate solution was obtained in the first evaluation.

We certainly learned from the ETS reports, particularly from their unique new Age Cohorts Design, and we probably had more time than they did for attempts to solve the more difficult problems that arose.

These same advantages were apparent when we examined the process whereby Sesame Street might cause learning, when we examined the kinds of children who might have benefitted from viewing, when we examined whether the magnitude of learning was sufficient to meet criteria of social or educational significance, and when we examined long-term effects and the unintended positive and negative effects. However, another advantage also became apparent in that we had access to a number of studies about Sesame Street that had been conducted after the ETS evaluations, and these allowed us to explore some questions in greater depth than would have been possible if we had only wanted to test if the ETS data supported the ETS conclusions. For example, the ETS team did not use nationally standardized achievement tests in their work, preferring to devise their own tests

that were closely tailored to the behavioral objectives of the show. Aptitude tests serve as indicator of transfer, and so we wanted to see if Sesame Street affected scores on tests like the Metropolitan. Minton[7] conducted a doctoral dissertation at Fordham in which she tested the effects of Sesame Street on the six Metropolitan subscales, and we were able to refer to her work.

In order to evaluate Sesame Street, we wanted to obtain data about the size and social composition of Sesame Street's national audience. Ball and Bogatz had not been commissioned by CTW to undertake this task, and so they limited themselves to statements about the percentage of their children in various categories who viewed the show for different amounts of time. Our focus on the national impact of Sesame Street, when coupled with the availability of viewing data from Neilsen, A. C. Harris, and Daniel Yankelovich, Inc., permitted us to answer the questions that we asked about the viewing audience, questions that the ETS team was not asked to ask.

Having established who viewed Sesame Street and what the gains from viewing were, we then asked about the dollar cost per year of reaching children who learned from Sesame Street, and we also asked whether the cost of reaching the advantaged was the same as reaching the disadvantaged. Once again, Ball and Bogatz were not commissioned to do this, and it is possible that they might not have been permitted to do so since they are not economists. Well, neither are we. But other persons have computed the costs of Sesame Street and it was possible to obtain, check, and re-compute some of their workings. Here again, the secondary analyst who sets out to do a comprehensive analysis has an advantage over the primary analyst who, first, will be less likely to gain permission to do a

comprehensive analysis and, second, will not have the time, self-attributed expertise, or desire to go into questions of cost.

These same points also hold for our last task in reevaluating Sesame Street which was an examination of the social values that the show might be promoting or hindering as well as a discussion of some policy recommendations. Primary evaluation researchers will typically not be allowed to conduct such analyses by program officials unless there is close supervision, for these are sensitive and subjective issues on which the evaluation researcher does not speak with the same authority as when he is analyzing data.

Utility of Secondary Evaluations.

We have previously noted that secondary analysts have advantages that come from access to the reports of the primary analysts, from access to the latest relevant research by others, from less tight deadlines, and from a greater independence in their relationship with the officials of the program being evaluated.

These factors can permit a reevaluation of the program that is more comprehensive than most primary evaluations or secondary evaluations that only set out to test whether the conclusions in a primary evaluation can be corroborated. Such comprehensiveness is important because it can highlight the implications of relationships between the findings from the different areas of evaluation. For example, we might find that Sesame Street teaches those it reaches. If we also find that it reaches more of the disadvantaged than th advantaged, then we begin to suspect that it teaches and is also narrowing the national gap in academic achievement. But if we find it reaches proportionately fewer of the disadvantaged, then we begin to suspect that it teaches those it reaches but in so doing it may widen the academic achievement gap.

In addition, secondary evaluations increase the chance that we can make statements about a program rather than statements about a program that are conditional upon the investigator. That is, if I find a functional relationship between X and Y that no one else can replicate, then the X - Y relationship is conditional upon me being the investigator. Alternatively, if I find the relationship and no one else has bothered to investigate it, all statements about X and Y are still conditional upon me being the investigator. To be more concrete, evaluation researchers have idiosyncracies that include how the data are initially controlled for quality, how respondents are omitted from the study or retained in it, how indices are constructed, how data are analyzed, how results are interpreted, and how the evaluation questions are posed. Until a second investigator with his unique and different idiosyncracies has examined the original data, we cannot begin to ascertain how much of the original evaluation reflects effects of the program and how much reflects the style of the original investigator.

What makes this point less academic is the possibility that the commercial nature of evaluation research may result in program administrators commissioning primary investigators whose style is biased towards demonstrating that a program is effective. This should not be taken as a reference to outright charlatanism. While charlatanism might benefit research firms in helping them gain future evaluation contracts from the agency whose program is being evaluated, it will not benefit them in the eyes of the organizations that fund the agencies to run the programs. There are countervailing forces at work here, for while the evaluator depends on the program, thereby increasing the chances of a pro-program bias, he also depends on the agencies funding the program, thereby minimizing bias because these agencies often

want as objective an evaluation as possible. In our opinion, systematic positive bias is more likely to enter a primary evaluation unwittingly rather than deliberatel . evaluator cannot totally forget the program piper who is calling the research tune. This is especially the case at crucial decision points where there is no objectively correct decision and where the nature of the decision can increase or decrease the chances of obtaining program effects.

How are primary evaluators chosen? In a study of scientists and the programmers they use, Inzirelli[8] found that the most important factor in the choice of programmers was the latter's identification with the project's goals, the next was the programmer's self-discipline, and the next was his competence. Obviously, we cannot extrapolate from choosing programmers to choosing evaluation researchers; nor can we generalize from a situation where there is probably little variability in programmer's competence to one where there may be greater variability in the competence of evaluation researchers. Nonetheless, it is worth investigating in the future whether program officials select evaluation researchers because the researchers identify with the program's goals and methods. If so, this selection process should exacerbate the other covert pressures on primary evaluation researchers to produce positive findings. The secondary evaluation will not be subject to these same pressures, of course.

A final role that secondary evaluations can play is to suggest new questions and new designs for the next primary evaluation of a program or of programs like the one being reevaluated. For example, our work on Sesame Street led to proposing some studies that are feasible and are needed if Sesame Street's national impact is to be assessed. One is a study of the way that schools do or do not capitalize upon Sesame Street. This study could be relatively easily accomplished because we were able

to demonstrate that at three of the seven ETS research sites there were randomly constituted groups of light and heavy Sesame Street viewers whose career through school could be charted in a way that will permit strong inference because of the original random assignment. Another study concerns the effects of viewing Sesame Street as opposed to the effects of being encouraged to view Sesame Street. ETS field personnel regularly visited the homes of encouraged children, urged them to view the show, and left behind buttons, picture books and the like. Encouragement to view was the experimental treatment in the ETS evaluations, and we can understand this in the historical context of Sesame Street's first year when it was feared that few children would watch the show. Most children in the nation view without encouragement, and we should like to be able to make as strong conclusions about viewing as we can about encouragement to view. A third study relates to the fact that there is only one national survey that explicitly includes data on the viewing of black children. The data come from the annual audience surveys of the Corporation for Public Broadcasting, but the surveys include only 60 black homes. It would be relatively simple to shift from the simple random sampling of homes to a stratified random sampling strategy which would increase the size of the sample of black homes and would increase our confidence that we know how blacks view Sesame Street nationally. Finally, we would hope that future primary evaluations of television programs that are equally available across the nation will examine whether these equally available opportunities are in fact equally used by all social groups. If they are not, the national implications of this might be pointed out. Secondary analysts are in a unique position to suggest important new evaluations of a program or of a particular kind of program. (For example, all the above remarks apply to evaluations of

the Electric Company as well as Sesame Street.) He is in this position because he knows the issues and the data, because he knows which important data are missing, because he has some idea about the cost of obtaining the additional data, and because his secondary evaluation can present the case for the additional information in a context where it is all the more likely to attract attention.

Putting the Advantages of Secondary Evaluations into Perspective.

We must be careful not to claim too much for secondary evaluations of programs. Though they may have advantages over primary evaluations or other forms of secondary evaluation they nonetheless have deficits. Let us illustrate the major ones.

The deadline problem that plagues primary researchers also applies to secondary evaluators. On the one hand, they may not want to spend too long analyzing other people's data given the current reinforcement contingencies in academe where, as we perceive it, the stress is on producing new data rather than reanalyzing old data. Also, the secondary evaluator needs funds, and the funding agency will impose deadlines and might not be open to extending these. Finally, if the secondary evaluation is of a current program, as it typically will be, the secondary evaluator will be under some pressure to feed his results into the decision-making process as soon as possible. Thus, he will not be able to sit back and ruminate like an academic monk in a cell so that every last i is dotted and every t crossed. He too will have to suffer some of the tribulations of hurried work.

Secondary evaluations are required so that we can distinguish statements about programs from statements about programs that are contingent upon the original evaluator. But even here two points are worth noting. First, a secondary analysis provides only one instance of corroboration so that,

strictly speaking, one can still only generalize to effects of programs that are contingent upon two evaluators or that are contingent upon any idiosyncracies that the primary and secondary evaluators share. Such generalization is better to have than not to have. But it is not dramatically confidence-inspiring. What makes it the less inspiring is that the secondary analyst may have compensated for some of the idiosyncracies of the primary researcher's data analysis but he will not have compensated for any of the idiosyncracies of his data collection (e.g. those that result in "Hawthorne Effects" or measures of outcomes that may have little correspondence with the construct whose name they are given and about which we want to make inferences). Thus, there will not be an independent test of whether the results can be replicated with a different data collection procedure or a different set of measures.

The second point concerns what happens when the results of the primary and secondary evaluator conflict. The secondary analyst is bound to have his own idiosyncracies of data analysis style, and these can be evoked to explain any discrepancies between his outcomes and those of the primary analyst. Critics might legitimately take exception to the way he aggregates data, the way he constructs indices, or the like. Sometimes, they will correctly point out mistakes, while at other times they will merely differ in their opinion. The point is that the secondary analyst is no God and that his work will be criticized. He will be at his weakest, perhaps, when he finds no differences where the original evaluator found differences, for we still lack adequate criteria for accepting the null hypothesis as opposed to failing to reject the null hypothesis. We are more confident of accepting the null hypothesis if 1) statistical power is high; 2) we are considering the maximal and minimal strengths of a particular treatment; 3) if there is unambiguous reason to expect a relationship between

a treatment and effect; and 4) if all the potential suppresants of the relationship have been controlled for. Yet how often do we have tests that meet these stringent criteria? I would venture that it is not often. Anyway, the point is that secondary evaluators have distinct analysis preferences that might impede them, and they tread on dangerous inferential ground with no-difference findings. This makes it inevitable that their work can be attacked and that the issue of the program's effectiveness can become embroiled in controversy.

We have previously touched upon one of the most serious drawbacks of secondary evaluations, the fact that they often cannot deal in adequate fashion with threats to the original evaluation that stem from the nature of the measures or the data collection process. Let us look at two examples of this, one from Sesame Street evaluations and the other from a critique of the Coleman Report. The evaluations of Sesame Street stressed the show's effects on the child's intellectual growth more than his cultural growth. Persons interested in cultural growth would have to content themselves with what is in the ETS reports or they would have to look elsewhere. Unfortunately, there is little other evidence of high quality about the show's effects on cultural growth, and so no general conclusions are possible about Sesame Street's effects in this area. Moreover, it is possible with Sesame Street as with probably the majority of other evaluations that all or some of the testers knew the experimental condition of the child they were testing or that the children and their parents were conscious of being in a study where the hypotheses were transparent and easy to comply with. Thus, tester biases or Hawthorne Effects could be invoked as alternative interpretations of any effects of Sesame Street. These interpretations would be difficult to refute with evidence that would

convince sceptics, and the secondary evaluator can do little more than note the threats and collect as much anecdotal evidence as possible about how testers went about their business and about how salient it was to children in different viewing groups that they were in a research study.

Bowles and Levin[9] criticized the Coleman Report for the large amount of missing data (only 59% of the high schools contacted responded to the survey, while many big cities like Boston, Chicago, and Los Angeles did not even permit the survey to be administered, and the nonresponses to items from returned questionnaires came disproportionately from poorer schools). There is little we can do abuut such missing data except to contend, as Jencks[10] did, that the magnitude of any resulting bias will be small. Unfortunately, judgements about the magnitude of bias are typically subjective and hardly helpful as the controversy between Campbell and Erlebacher[11] and Evans and Schiller[12] over Headstart makes clear. Next, Coleman's major analyses of the "effects" of expenditures on achievement involved school districts as the unit of analysis rather than schools. The effect of this was to lump together the schools in each district that spent most and least per pupil and this resulted in a variability between school districts that was presumably less than the variability between schools within school districts. Thus, according to this procedure Chicago would have been assigned a single per-pupil expenditure score (if it had been in the study!) and any difference between Chicago schools would have been ignored. In this situation one cannot make any generalizations beyond the restricted range of school expenditures represented by a biased sample of school districts, and one probably could not choose out extreme comparisons of the best- and worst-equipped schools. The secondary analyst could only increase the variability in per-pupil expenditure if he could

disaggregate the data. It will be sometimes possible to do this by indirect assessment methods. Armor[13] assumed that the major expenditure was for teacher salaries and so he averaged the salaries for the teachers at each school who reported their salaries, then he multiplied this average by the principal's estimate of the total number of teachers, and he finally divided the resulting figure by the principal's estimate of the student enrollment. Such ingenuity can sometimes lead to estimates of how the disaggregated data would have been, and so the secondary analyst is not at a total loss when the data are not in the form he needs. But he will often be at a loss.

The final problem with secondary evaluations concerns access to the raw data. Some investigators feel free to pass on copies of their data to anyone who asks; others are likely to pass them on to friends or respected professionals; others may pass them on out of fear that not passing them on will be seen as unprofessional; others may pass them on when high prestige organizations intervene and use their "clout"; others may pass them on because they do not own the data and the owner (e.g. an office in the Federal government) simply takes a copy of the data and passes it on; while others may pass on the data wher they learn that the secondary analyst is interested in questions that the program administrators do not feel to be sensitive. We know of instances where each of these factors was probably at play. What is important to note is that program officials, if they want to restrict or delay access to the data, can typically do so. Even when they do not own the data, they can stall for time in the hope that the secondary analyst will lose interest or in the hope that the issue of the program's success or failure has decreased in salience. Even legal action based on the Freedom of Information Act is likely to be time-consuming, and the investigator's only hope of speedy access to publicly

owned recalcitrant data is to have the program funding agency come right out and demand the data. But this is not likely to happen, especially before a final report has been issued. When data are privately owned the situation is even more difficult for the secondary analyst whose proposed work is not appreciated by program authorities. As far as we know, he has little recourse except to invoke norms of an organization's public accountability for publicly available reports based on data.

The foregoing limits to secondary evaluations should help us understand that their advantages, while real, are not unlimited. What we need are ways of increasing the utility of secondary evaluations and of eliminating the obstacles that stand in their path. The next section deals in gross detail with some ways of achieving these goals.

Guidelines for Improving Secondary Evaluations.

The access problem could be solved if it were the policy of Federal and foundation authorities to request that a copy of the complete raw data, of all research reports based on the data, and of all instruments used in measurement, be deposited in some well-known location, like the Roper Institute at Williams College, or ISR at Michigan, or anywhere similar. The data would have to be deposited at the same time a final report is submitted, and would be available to all persons who are willing to pay the costs of buying tapes, research reports, Xerox copies of measurement instruments, and a small handling charge. This procedure would have positive consequences other than facilitating access. First, it would remove the secondary analyst from all contact with program personnel and from most contact with primary evaluators; second, it would make the data available for examining theoretical and practical issues to which the data are relevant and which may have nothing to do with the issues in the primary evaluation.

One difficulty arises with this procedure if secondary analysts need information about features of a primary evaluation that are not mentioned in detail in reports (e.g. data-collection or tester-training procedures). We would want secondary analysts to have such information and we would not want primary analysts to be bothered replying to all enquiries. Hence, we might have to develop a procedure whereby a standard questionnaire about the most frequent of such details is given to primary evaluators and their replies are made available to secondary evaluators along with the data. Requests that cannot be accomodated by such a scheme are likely to be few and probably could be obtained without too much inconvenience by writing to the primary analyst. Another difficulty arises about what constitutes a final report, especially with multiple-wave panel studies where defensive program or primary evaluation personnel might argue that any release of the data before a report on the very last wave could lead to erroneous interpretations. In such a situation, one might want to specify that secondary analysts should have the right to gain access to any data on which progress reports are based that draw any conclusions about the effectiveness of a program. A further difficulty arises because of the privacy issue. Sometimes, secondary analysts may want to mesh one set of data with another, and this requires the identification of individual respondents. However, schemes could be set up so that neutral parties (e.g. the data depository center) could do the meshing and no one who conducts a secondary evaluation would be able to link the data with anything but code numbers.

A quite different strategy might be appropriate for solving the problem that a single secondary analysis will inevitably reflect the stylistic idiosyncracies of the secondary evaluator. For the primary evaluation of important programs it might be better to move to a

decentralized data analysis model such as the one that will be used for analyzing the results of Project Prime, an in-house research project of the Bureau for the Education of the Handicapped. Here, one agency had responsibility for the research design and data collection, and another for data reduction and editing. But the edited data will be analyzed by at least three separate teams of data analysts. This will reduce the risk that is involved when a single agency takes responsibility for analyzing a large body of important data, and the quality of the evaluation is in large measure dependent on one or two competent but fallible persons. If a decentralized data analysis strategy were followed in primary evaluations, there would probably be less need for secondary analysts. But since most programs will not merit the relatively more expensive procedure of decentralized analysis, and since some programs may begin modestly and be escalated to "success" in part because of the results of a single primary evaluation, there will obviously be a continuing need for secondary evaluations. There is no reason why secondary evaluations should not follow the model of decentralized data analysis, and Federal or foundation authorities might sometimes want to consider providing funds for several different secondary evaluations of important national programs. After all, the analysis of edited data is one of the least expensive parts of an evaluation.

An apparent problem arises in simultaneously advocating comprehensive secondary analysis and the speedy publication of the results of secondary analysis. Speedy publication requires early deadlines and early deadlines inspire both carelessness and evaluations of limited scope. Perhaps this is why most secondary evaluations in the past have concerned themselves mostly with the questions of whether a program causes statistically

significant gains among recipients of different social or racial groups, and they have not taken the larger perspective that we have previously outlined and that we took from Suchman's fine book called Evaluative Research. One way of getting around this problem would be to recognize the interdisciplinary nature of evaluation research where, if we look merely at the luminaries in the field, we have statisticians, political scientists, economists, sociologists, educators and psychologists, and where we have persons in government, in foundations, in academe, and in the private sector. Would it be too much to hope that at some time we might have three or four centers of Evaluation Research throughout the country where primary and secondary analyses could be conducted and where there would be sociologists on hand who know how to measure socio-economic status (it is staggering how many educators do not know this!) and who can advise about the value implications of findings, where there would be economists willing to conduct cost-benefit analyses, where there would be psychologists willing to create strong research designs, and where measurement experts would be willing to devise new measures or advise on the suitability of extant measures or construct new indices, and where applied statisticians would be willing to help with the data analysis? One man can do many of these things; but he will be hard pressed to do them all especially if he has deadlines to meet.

The important point is that there be interdisciplinary teams, irrespective of how they are organized. To be sure, there are difficulties with this; it is difficult enough to get people to cooperate when they come from a common discipline, let alone from different disciplines; moreover, large organizations often require what seems to be an inordinately long time for internal organization and this prevents people from getting ahead with the primary or secondary evaluation. After all, the pressure

of deadlines on any one man is only eased, and his work improved, if others free him to devote more of his time to his particular research tasks rather than to organizational tasks. Another difficulty is that we may not all agree about the desirability of organizations that need large budgets, charge high overhead, and are dependent on contracts for their survival. This creates some of the very problems that commercial profit and nonprofit agencies face and that give rise to the need for secondary evaluations in the first place. I would suggest, therefore, that interdisciplinary evaluation agencies be attached to universities with a reputation for evaluation research and that, when contracts arise, the cooperation of experts from the parent university and other universities in the neighborhood could be purchased so that, for however long it takes, they could be working full-time or for a large portion of their time on the primary or secondary evaluation project. The Harvard Seminar on Equality of Educational Opportunity had some such interdisciplinary features, though the seminar members did not always reveal as much of their respective uniquenesses as I would have liked (because most concentrated on testing the same issues as were attacked in the Coleman Report and they did not focus as much on the implications of findings or on comparative analysis with other studies of achievement and school resources). We are not suggesting that the brief organizational outline above is the only one possible for evaluations. However, we are suggesting that there is a national need for viable organizations that permit comprehensive interdisciplinary primary and secondary evaluations and that allow the analysis of edited data to be conducted at several places. Of course, this kind of expensive work is only called for when nationally important studies are involved. But the number of these is increasing as the national commitment to "planned variations" or experimentally planned reforms increases.