DOCUMENT RESUME

ED 195 571 TM 810 018

AUTHOR Meyers, James M.
TITLE Integrity Versus Pragmatism in Large Scale Evaluation: Planning the National 4-H Evaluation.
PUB DATE Apr 80
NOTE 32p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, Ma, April 7-11, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Accountability: Economic Factors; *Evaluation Methods: *National Programs: *Program Development: Program Evaluation: Social Influences; *Youth Programs
IDENTIFIERS 4 H Clubs: Food and Agriculture Act 1977

ABSTRACT

In 1977 Congress mandated an assessment of the "economic and social consequences," of the nation's 4-H programs. The evaluation problem posed involved the identification of all social and economic effects caused by some facet of a very large, nationally decentralized youth education program. The design utilized drew on exploratory research methods to focus on inferred, attributed, and demonstrated consequences at three orders of occurence. This mandate sparked conflicting concerns between sponsoring and administering stakeholders for "integrity" and for "pragmatic necessity" in evaluation design. In the course of dealing with these conflicts, several potentially critical differences in large and small scale evaluation situations were identified. Some of the differences suggested included: (1) nature of the evaluation question, where large scale evaluations (LSE) tend to assess underlying funding policy or rationales and smaller scale evaluations (SSE) tend to assess program operations: (2) program variance, which increases with scale: (3) data management and collection, where LSE are more costly than SSE: (4) effect of time and resource constraints, which are more likely to effect design and methodological choices in LSE: and (5) stakeholders' attitudes and behaviors, which seem to increase with fear, suspicions, and controversy as scale increases. (Author/RL)

INTEGRITY VERSUS PRAGMATISM IN LARGE SCALE EVALUATION PLANNING THE NATIONAL 4-H EVALUATION

By

James M. Meyers
U.S. Department of Agriculture
Science and Education Administration

April 1980

This paper has been prepared for presentation at the annual meeting of the American Education Research Association, April 1980.

Support for the research on which this paper is based has been provided by the USDA effort to respond to the requirements of Title XIV, Section 1459, of the Food and Agriculture Act of 1977. However, the views expressed in this paper do not necessarily represent the official views of the U.S. Department of Agriculture.

INTEGRITY VERSUS PRAGMATISM IN LARGE SCALE EVALUATION PLANNING THE NATIONAL 4-H EVALUATION

How would you go about convincing reasonable, but skeptical third parties of the value of your total educational organization? How would your staff respond to the question, not of how well local entities accomplish their goals, but of the resulting social and economic consequences of your entire organization and program? If, like the staff of the 4-H program, you would choose to rely on a past-proven formula of describing broad program goals (hopes really) for participants, backed up with input measures of the extent and efficiency of your programs capacity to reach participants and volumes of successful case examples and participant testimony, you may be surprised to find that evaluations predicated on the inherent goodness and value of providing educational programs are likely to find their design integrity challenged by budget-minded policymakers.

Integrity and pragmatism in evaluation design are hardly new concerns in evaluation literature. Further, integrity and pragmatism are not necessarily conceptually opposed concerns. However, they often seemed so in planning a national scale evaluation of the 4-H program and, in reflecting on the past decades methodological debates in evaluation literature as experimental models have been argued against decision models, etc., conflict between these concerns does not seem surprising.

The objectives for this paper are threefold. First, to describe the experience of 4-H with what may prove to be a more prevalent type of evaluation question, focused in a national scale evaluation. Second, to explore how tensions for both design integrity and practicality were manifested and affected a large scale evaluation. Third, to describe the design 4-H employed to respond to a Congressional evaluation mandate.

Four-H is one of five major programs conducted by the U. S. Department of Agriculture, the Nation's land grant universities, and local county governments cooperating in the sponsorship of the Cooperative Extension Service. 1/ Four-H is best known as an informal, quasi-vocational program for farm youth. Nearly everyone is familiar with 4-H boys and girls with their cakes and cows at the county fair. Like most social and educational programs, however, 4-H has become much more than its popular stereotype. Modern 4-H encompasses subjects ranging from dog care and training and bicycling, to career development and the study of human nutrition. Four-H enrolls inner-city youth as well as rural youth. It has grown from an almost compensatory educational program for rural youth only to a general youth development effort for all youth. Four-H involves some 5 million youth, 550,000 volunteer adults, and about 8,500 paid staff nationally. Annually, 4-H receives approximately $150 million in public funds and about $2.2 million from private sources. Like all Cooperative Extension programs, 4-H was established as a means of extending useful information and technology deriving from university-based research to local communities.

In its major farm bill of 1977, the Congress in one very brief paragraph charged the Secretary of Agriculture to, "transmit to Congress, not later than March 31, 1979, an evaluation of the economic and social consequences of the programs of the Extension Service ..." 2/ The term consequences is not a common one in evaluation literature, most likely because if not modified so as to be specific to program goals or some other interest, it really includes all effects attributable to the existence of a program.

The context of educational evaluation has been rapidly changing as both researchers and administrators have gained experience and expertise in the conduct and use of evaluations. Early evaluation charges tended to ask evaluators to determine how well programs were meeting established objectives. More re-

cently, as program objectives have come to be understood to be only a part of outcomes of interest, evaluators have been directed to determine how well programs meet the needs of recipients, regardless of objectives. It would seem that now, by the charge for this evaluation, that yet another level of evaluation mandates may be emerging. The focus on social and economic consequences directs the evaluators to determine all social and economic effects, unlimited by program objectives or recipient needs.

This mandate presented a charge that went far beyond 4-H's evaluation experience. Many 4-H program and research staff stated that it was practically unmeasurable and was inappropriate and unreasonable in its implication that 4-H should be held accountable for its participants' social and economic behavior. The resulting internal debate and dilemma has touched such deeply held feelings that there is some reason to be concerned that the report of this effort may not be well accepted nor be directly followed up on. Nonetheless, we believe that the evaluation question as framed is not only reasonable and appropriate coming from those who fund the program, but may also be most relevant to national or other broad scales of oversight.

The implications of scale of inquiry in evaluation design has not been treated with over much attention in evaluation literature. For the most part, evaluation literature has featured smaller scale evaluations. No doubt because those are the most prevalent. However, it also seems tenable that we have generally assumed that large scale evaluations should be methodological extensions of smaller scale work. That is, that large scale evaluations should simply involve increasing and redistributing the data base, but be otherwise similar to smaller scale inquiries.

In fact, however, the problems of large scale evaluation may not be so similar to those of the smaller scale. The often inflammatory debates that seem to characteristically follow large scale program evaluations should perhaps have signaled that there may be significantly different concerns and issues related to scale of inquiry. There seems seldom a national or regional scale evaluation published that is not well taken to task, if not completely repudiated, on methodological grounds. 3/ Of course it may be that those who design and conduct large scale evaluations are not as competent as those who oversee smaller scale efforts. However, it seems more likely that significant differences in scale of inquiry may be associated with significantly different evaluation questions, situations, and therefore in expectations. This paper proposes several possible differences and to the extent any can be definitely related to scale of evaluation inquiry, there is need for reconsideration of methods and expectations for large scale evaluations.

Four-H would be difficult to evaluate on the national scale under any circumstances, since there is in effect no one 4-H program nationally. Local extension agents (in most counties in the U. S.) are employees of their state's land grant university, by whom the programs are largely managed. Four-H programs are not subject to national centralized direction. Rather, local program staff are free to adapt general educational programs to local situations, since local participants do so only voluntarily as they perceive the program's utility and value. In the case of 4-H, this is particularly critical since many efforts are predicated on high rates of adult participation as volunteer staff. There is considerable potential for variability in 4-H programs not only from state to state, but from community to community within states as well as agents modify their practices to juxtapose with local preferences. As a result, one of the

possible differences associated with scale is the compounded programmatic complexity (i.e., increased variations in organizational practice and design) that larger scale evaluations must contend with.

THE QUESTION OF INTEREST

As soon as efforts to better define this evaluation's key questions and choose an appropriate design began, apparently conflicting concerns for both the integrity and the pragmatic necessities of the study surfaced. The terms integrity and pragmatism were in fact used in internal discussions (sometimes debates), but they are used here in even broader context to represent sets of real fears and objectives which have been expressed.

Emphasis on pragmatism in design and effort tended to come from 4-H program practitioners and administrators, who, having usually invested considerable time and effort (to say little of self-identification) in securing the program's daily survival and well-being, had understandable fear that "uncaring" evaluators might overlook or discount hard-won qualitative values in the effort to apply sophisticated methodologies. Thus, there were fairly consistent admonitions to avoid emulation of "research" or experimental methods and to focus instead on pragmatic methods. Emphasis on the integrity of the design and effort tended to come from oversight-minded representatives of other agencies (as well as from the research community). These were groups responsible for the Federal sponsors input, who wanted to assess the value to them of what they saw as their investment. Being generally fairly experienced and sophisticated with program evaluations, they rightly pointed out that while strictly correlational or perceptual data are better than no data, they are not satisfactory evidence of causal relationship to the program. From these parties were heard consistent admonitions to avoid program assumptions about outcomes and causes and to focus instead on developing a design which would have maximum external credibility.

Relations between extension's Federal and state partners have always been characterized by a mild level of tension, although organizational rhetoric maintains otherwise. Of the many probable causes for this tension, two stand out. First, the states are jealous of their autonomy and have some fear of Federal dictation. Second, administrative agency officers sometimes have trouble understanding why extension's numerous field agents cannot be marshalled to sell administration programs at the local level. Complicating both of these concerns is the fact that agents' local effectiveness is at least in part predicated on being perceived as advocates for local client interests rather than either state or Federal interests.

Although not strictly related to the evaluations scale, it is important to note that this evaluation focused on a question put by major program investors, from the perspective of their reasons for investing. There is no question that the identity of the questioners, their expressed attitude of suspicion toward program staff, and the fiscal and administrative accountability nature of the question exacerbated usually mild tensions. While it is clearly possible for the question of "program consequences" to be applied between similar actors in smaller scale evaluations, it does seem a much more likely circumstance for large scale evaluations. Such a situational difference could grow from the probable identity and interests of evaluation sponsors at each scale. Smaller scale evaluations are likely to be initiated by program administrators, who will more likely have interest in whether or not the program is efficiently performing as designed. Thus, smaller scale evaluations could be hypothesized to more likely focus on goal attainment, operational practice, etc. Larger scale evaluations (i.e., national, regional or state scale), are expensive and complex and

seem more likely to be initiated (as the 4-H study was) by fiscal sponsors, whose interest is in assessing the value of an investment. Thus, large scale evaluations could be hypothesized to more likely focus on total program impact and worth through charges such as that given 4-H.

While one might otherwise entertain some doubt over the mutuality of legislative and executive branch interests, Congress had given the charge to the Secretary of Agriculture and was not disposed to elaborate further on its interests. Both USDA and OMB proved to have fairly specific interests. They were interested in "hard" data on the full range of consequences. While they did not eschew qualitative or purely correlational data, they wanted acceptable-to-themselves accounting of consequences and the programs' causal role. The interest in 4-H's causal role was less that of scientific curiosity than what came to be termed "budget-based" interest. That is, information was wanted which would facilitate budget decisions about the program's future. Essentially, these program sponsors wanted an accounting of program consequences which they could then assign value to according to their own priorities. Further, they were interested in cause for two reasons. First, because 4-H had historically only provided "soft" data (e.g., successful case examples) to them. Secondly, because they wanted assurance that any effects they valued were really products of their investment in 4-H and not multipally caused effects which might well accrue with lesser or no Federal support.

Staff efforts to define consequences in some more definite terms yielded an outline which caused most stakeholders to react. Consequences were defined rather literally to mean events or effects produced by some preceding action or causal contribution of the 4-H program. Within such a definition the simple existence of 4-H programs would produce consequences, not to speak of proactive

or reactive program activities. The term consequence was further defined to refer to the full realm of program effects, not simply intended outcomes. Included were consequences which might accrue more immediately and those which might accrue at deferred times and places. Also included were those which might accrue to individuals or smaller groups and those which might accrue to populations or major institutions. This definition was only modified by the terms social and economic, which in fact, did little to narrow the universe of consequences of interest. Evaluation literature seemed to hold little in the way of examples or guidance for such an effort, except the encouragement and precedent reasoning in Michael Scriven's development of the idea of goal-free evaluation. 4/ The field of consequences was ordered as follows to help focus interest and ensure attention to varied major areas of potential effect.

> First order consequences--effects accruing to indivudals and small groups relatively immediately, e.g., participation, acquisition of information, skills, attitudes, etc.
>
> Second order consequences--effects accruing to individuals and small groups across time and space (i.e., in deferred settings), e.g., changes in behavior, changes in income, health, quality or style of life, etc.
>
> Third order consequences--effects accruing to community or regional scale groups and/or social or economic institutions as an aggregate effect of first and second order consequences, e.g., changes in community practices or capacities, changes in income distribution, health indices, social norms, laws, etc.

It was additionally noted that: (a) Consequences might accrue to both direct program participants and to nonparticipants; (b) consequences might be

directly and solely caused by the program or they might result from the interaction of multiple forces, of which the program might be a necessary but insufficient part; and (c) the relationship between consequences at each of these levels would not necessarily be a direct cause and effect linkage.

The publication of this view of consequences raised immediate objections from 4-H program staff at both Federal and state levels. Generally, they labeled it "overly research oriented" and beyond the realistic constraints of time, money, and contemporary methodology. USDA and OMB representatives reacted more favorably, supporting the attention to cause and the open-ended, or relatively nondirected, form of inquiry. They did indicate that their primary interests lay with what had been termed third order effects, these being more consonant with national scale planning scope.

Staff also made an effort to quickly review available data from 4-H program studies and evaluations. There was, in fact, very little program outcome data that went beyond what had been termed first order consequences. Furthermore, although 4-H was over 60 years in existence, there were relatively few outcome studies at all. Three factors, which are probably not uncommon to similar programs, seemed to have contributed to this paucity of outcome information. First, 4-H had enjoyed some 60 years of continuous, relatively unchallenged Federal support, so that little need for nationally usable outcome data had been perceived. Second, 4-H still clung to the pre-war-on-poverty assumption that educational programs were inherently good. That is, that the dissemination of knowledge was an acceptable end unto itself. This particular view was further complicated by a seeming extension of thought about individual rights and academic freedom so that there was a program rhetoric that held that not only should the program not be held accountable for what were termed third order con-

sequences, but that it would be counter to the concept of individual freedom to plan programs to deliver specific third order effects.

Thirdly, because 4-H is largely focused on developing and maintaining voluntary participation, program staff have traditionally emphasized studies of input-enrollment efficiency and the program's ability to reach increasing numbers of participants. This emphasis is underlain by the belief that only persons satisfied with the value and utility of outcomes would become or remain participants voluntarily. The result of these factors was that 4-H, no doubt like many voluntary, client-focused educational programs has invested nearly all past accountability effort in recording numbers of participants reached, the subject information with which they had been reached, and gathering widely spread success-story case examples. Furthermore, the program's unpreparedness to respond to the type of outcome evaluation demand posed here extended beyond the lack of organized and useful information. Program staff, with years of successful experience justifying the program with input, enrollment, and loosely gathered qualitative narratives, were neither convinced of the need for nor the methodological considerations of a different approach.

INTEGRITY VERSUS PRAGMATISM

As evaluation staff shared the results of these three early inquiries with 4-H program stakeholders several variants of both pragmatic and integrity positions were put forward. As noted earlier, 4-H program staff for the most part, put their methodological preferences forth as pragmatic. Three types of alternatives were generally suggested by these proponents. First, were suggestons of a design built around correlative measures of traditionally held program values. Specific examples suggested included comparisons of 4-H participants and nonparticipants on measures such as rate of contact with juvenile authorities or levels of education, income or social achievements as adults. A frequently

suggested variation of decision-focused evaluation was extension of the assumption of the inherent value of dissemination of information relative to selected national needs or priorities and the subsequent measurement of 4-H input investment and rates of participation in "need" areas.

Probably the most compelling of the pragmatist-based design suggestion was put forward by program staff arguing that consequences are individually unique and that exact identification and scientific proof of cause with existing methodologies were simply impractical. These proponents argued that,

> ... the program's value doesn't lie in whether or not it's produced huge and dramatic consequences. Four-H's value lies in the individual things that individual members and families get from it--important to the individual and to a few others around that individual, but not in terms of major public policymaking consequences.

Design suggestions from this group, put heavy, if not exclusive, emphasis on participant perception of benefit surveys. Such a position has natural and deep roots in most voluntary education programs, which have a strong focus on orienting program design, publicity, reinforcement, etc., to the participant population's perceptions of need and satisfaction.

Arguments for pragmatic focus of design all tended to share some degree of rejection of the right or appropriateness of government policymakers to hold an educational program accountable for facilitating broad social or economic interventions. In rejecting the practical or appropriate measurability of third order consequences with any investigation of the program's causal role, pragmatists tended to urge a design built on aggregation or expansion of the program's traditional smaller scale methods. Somewhat disturbingly, many of these positions also contained the veiled threat that any attempt to treat with 4-H's causal role at the third order would be met with serious critiques based on the proposi-

tion that cause can only be addressed with experimental or quasi-experimental designs.

USDA, OMB, and other representatives of oversight groups continued to voice integrity-based methodological preferences. Here too, three general types of design suggestions predominated. The most extreme of these argued that if the program's causal contribution could not be proven in nationally generalizable terms, then no "consequences" could be claimed at all. Design suggestions were almost classically scientific, moving from interview surveys to establish consequences variables to their experimental, nationally generalizable measurement. The other extreme among integrity focused views was not specific as to preference among correlative, experimental or other strategies as long as data could be quantified and an unequivocal determination of the program's overall success or failure in terms of a specific objective was stated. One party supporting this view went so far as to share a set of "exemplary" previous evaluations, of which no common pattern of design or methodology could be discerned, all of which, however, utilized quantified data to predicate holistic judgments of program value. The majority of oversight interests took a much more centrist integrity-based position. This was essentially that a description of program consequences, including objectively plausible explanation of the program's causal role, was needed to assess 4-H's role among similar educational efforts and its role in social and economic change. These proponents tended to eschew the need for classically experimental proof of cause. Instead, they suggested that a rational, testable explanation of cause supported by empirical or other objectively acceptable data would suffice. Designs suggested tended toward complex combinations utilizing accepted theory, qualitative, and quantitative data in an integrated inquiry.

Integrity-based views tended to strongly emphasize interest in the impact of the program in contributing to social and economic conditions as a function of government investment. Another strong theme among all integrity-based views was that internal, programmatic assumptions about cause or consequences were not acceptable as a basis for evaluation design. The 4-H programs oft-repeated assertions that program participation led to social and economic success was simply not acceptable without some form of objective validation. The dilemma over integrity and pragmatism developed largely into the differing views held by program staff and program sponsors over for what and how 4-H should appropriately be held accountable. As the issue of trust or credibility between program practitioners and one set of sponsoring partners emerged, it became self-reinforcing as program staff tended to take on more self-protective stances and deeper suspicions were triggered on the part of government policy analysts.

There seems little reason to doubt that this deb-te between integrity and pragmatism in evaluation design is more attributable to the very different views and expectations held for educational programs by policymakers and analysts and by educators than it is to the scale of inquiry itself. However, as has been suggested earlier, it does seem reasonable to expect larger scale evaluations to involve confrontations between these stakeholders more often than would smaller scale evaluations which may be one reason for the seeming proclivity of large scale evaluations to become embroiled in controversy.

The problem of sponsor-performer suspicion in 4-H's case seems to have had its roots in a gradual devolution of direct negotiative relations between the program's sponsoring partners (Federal, State, and local Governments) and the gradual assumption by national program staff of the task of keeping each partner sold. As dialogues on evaluation focus and methods proceeded it became apparent that even basic expectations about the utility of the 4-H program

(whether as a tool for social change or as a social maintenance service system) were not only not agreed upon, but were not even clearly understood. It became clear that this evaluation mandate was a sign that basic historical agreements and understandings between the state and Federal partners in the extension system could no longer be presumed upon. In many ways, national program staff in trying to serve what they perceived to be program needs rather than mediating stakeholders' interests, had come to stand between program practitioners and the various governmental and public interests which co-sponsor 4-H for their own reasons.

STUDY DESIGN

At this point, two developments occurred which clinched the adoption of what has been described as the integrity-based view. First, was the evolving realization by some program administrators that the program's accountability credibility with a major funding sponsor was truly failing and needed restoration. Extension administrators most in contact with the evaluation began to give their support for integrity-based design. These were not a majority, but they provided important evidence of the program's growing need to develop new evaluation emphases. The second development concerned discussions with senior OMB staff. Evaluation staff pointed out that their assessment of the evaluation situation and current state-of-the-art educational evaluation methodology made unlikely very definite causal assessment of the full universe of specific consequences. In effect, that some trade-off between external credibility, in terms of exploring the full range of potential consequences, and degree of consequence specificity was necessary. Faced with the choice of more specific measurement of selected variables or a wider search of potential consequences with less specificity but improved confidence in variable identification, OMB staff opted in favor of the latter as most consistent with their value for project integrity. While not

satisfied with such a trade-off, they indicated that they did not expect to take any imminent budget action as a result of this evaluation, and would prefer to see a (to them) credible ongoing effort established which would be expected to address specific consequences to the third order in the future.

In making this concession, however, Federal sponsors asked for increased descriptive information on the program and its management. A series of policy-based questions 5/ were articulated to be addressed in developing program descriptions and data on consequences.

These included questions such as: (1) How does 4-H determine program priorities, (2) how responsive and effective has 4-H been in identifying and serving new clients, (3) what alternative educational methods are used, (4) to what extent has 4-H been responsive to national needs and objectives? These questions were treated, but with less emphasis than the mandated charge to identify social and economic consequences.

Having settled that the question of interest was to identify as broad a universe of social and economic consequences of 4-H programs as possible, with at least some explanation of the programs causal role and mechanism, the evaluation team quickly settled on an exploratory research design which would collect and organize data of various types, from various sources in such a manner that a framework of probable consequences could be developed for which some at least potentially testable link to program experience could be identified.

Typically, evaluators might begin such an inquiry from the perspective of the programs stated objectives. However, undoubtedly like many educational programs, the objectives of 4-H are so broadly stated as to require substantial translation before they could be measured.

Consider this example: "To help young people develop inquiring minds, an eagerness to learn, and the ability to apply science and technology." In the first place, it is not readily apparent how to measure such an objective. Inquiring minds can express themselves in myriad ways. However, in evaluation, as opposed to basic research, more subjective measures can be considered, so for example, one might ask participants their own opinions, or seek to survey scientists and technologists and see how many were in 4-H, etc. Setting aside for the moment concern for measures which are solely correlative and give no definite clues as to cause, this approach would quickly narrow the scope of the inquiry to only those consequences previously noted by 4-H staff.

To have done so would have been contrary to the accepted position regarding integrity which specifically indicated that the inquiry not depend on in-house assumptions about outcome and cause. To have substituted some other informed, but largely subjective, judgment as to what outcomes should be selected for measure would have simply transferred the burden of satisfying external concerns about bias. The best, most objective measures available are, after all, only as good as the method used in determining what should be measured.

Thus, design choices proceeded from the assumption that 4-H had not yet identified all, nor even necessarily the most important, consequences for measurement. Exploratory research designs begin, in effect, with the assumption that we are not satisfied with our capacity to identify the variables of interest (social and economic consequences) for empirical measurement and assessment. As noted earlier, there is precedent for such an approach in Scriven's "goal-free" evaluation idea.

Essentially, the exploratory design decided on to generate the universe of potential, if not probable, consequences was to involve three generically different types of data. In increasing order of confidence, these were as follows:

A. Attributed consequences (i.e., effects perceived and attributed, but not causally accounted for or empirically demonstrated).

Attributed consequences were to be identified from three different sources (keeping in mind that attributions are here distinguished from demonstrated conclusions).

1. A review of attributions made in popular literature.
2. A review of research literature.
3. A review of participants and others' perceptions.

B. Inferred consequences (i.e., effects inferred from the comparison of accepted theory or studies of comparable units so that cause and/or demonstration is accounted for). Inferred consequences were to be generated in four separate theoretical studies. Two conducted by theorists outside the USDA-land grant university network and two by researchers who work with the 4-H program. The charge to these theorists was to review provided descriptive material on 4-H and how it operates, to review actual field operations, and to extrapolate as broad a universe as feasible of social and economic consequences citing the basis of currently accepted educational and other relevant theory.

C. Demonstrated consequences (i.e., effects demonstrated empirically in studies of 4-H programs).

1. Ninety-one selected studies of various 4-H programs, all of which involved specific outcomes were submitted to two external contractors for appraisal. These data represented the best, most recent studies of 4-H programs.

In order to objectively ground the study, extra effort was expended to collect a wide range of descriptive information about the 4-H program and how it

operated. The effort at thoroughness in developing descriptive data was expended over concerns for brevity and neatness of outcome focus for several reasons. First, if the programs' self-perceptions were not to become the basis for the inquiry, descriptive data is needed as a basis for inferring consequences. Additionally, general program description would serve as a backdrop to help identify consequences which might be missed with a data focused (or limited) strategy, and it would serve as a store house of information potentially applicable to consequences identified late in the study. Finally, any attempt to impute value to consequence findings should be done with knowledge of context and situation.

Within the descriptive effort, several special studies were initiated to address policy questions known to be of special interest and of less availability in existing data. These included studies of:

A. Decision processes.

B. A profile of 4-H alumni.

C. Data on 4-H participant's family income.

D. Data on public awareness of 4-H.

E. Data on the status and needs of 4-H age youth.

F. Data on other national youth organizations.

G. Staff surveys on administrative organization, research, funding, membership, etc.

H. Numerous descriptive narratives of program components.

Because 4-H is not subject to nationally centralized direction, but rather provides for locally modified practices to juxtapose with local preferences, the next concern was with how to deal with 4-H's potential for widely variable effects. More than once, program staff had negated evaluation requests by asserting that variance in effects was so great as to preclude useful measurement, except on a group-by-group or community-by-community basis. At the same time, it was recog-

nized that some means of developing units for analysis had to be found which provided less variance within them than between them.

Descriptive data suggested that there were widely used patterns in 4-H's experiential and curricular offerings.

MacDonald and Clark in summarizing curricula analysis literature observe that "... separating objectives from curricula materials and ... from instructional treatments is extremely difficult." Therefore, they assert, "the smallest viable empirical research unit with useful explanatory power would seem to be what is called a treatment ..." which includes those factors above. 6/

Traditionally, 4-H, like many other education programs, has recorded and analyzed program input according to categories of subject content involved. However, 4-H tends to extend most of its subject content via a wide range of experiential participation units (i.e., clubs, television, camps, etc.). Apart from the suspicion that the experiential differences in these various groupings might generate more variance in consequences than subject content alone, the use of subject alone falls far short of MacDonald and Clark's notion of educational treatment.

Seven major differentiated treatment units were identified. Four-H staffs in all 50 states were surveyed to see if they concurred with these categories as defined and if they could account for all of their participants using them. All of these categories were not new. Types of participation had been differentiated in enrolling youth, but no prior effort had been made to define and explain them formally. Surveys indicated nearly 100 percent ability to apply the units of participation as they had been defined.

There was now at least a plausible and testable method for dealing with variance of effects as well as a means of ordering descriptive information to provide a link to theories developed in other contexts and other data on potential consequences.

As these data became available, they were to be ordered and compared to see if a set of generic consequence categories (or a conceptual frame) could be developed. This first attempt to summarize consequences was to be program-wide and general enough to contain the wide ranges in individual variance involved. It should be noted that since the linkage of first to second to third order consequences becomes increasingly subject to interacting, individual factors, the variance in consequences that must be allowed for will increase as well. Thus, it was expected that first order consequences would prove most specifiable, with the second and third orders becoming increasingly broad. It was for this same reason that the second and third orders were expected to prove the least studied and the most problematic to predict.

At this point, the original schedule called for issuing the report to Congress. That report was outlined as follows:

A. Historical analysis--to give some dimension to how 4-H evolved in relation to social and economic change. What 4-H is today, is at least in part due to what it has been and the relationships and expectations built in its 60-year plus history.

B. Description of the program--to include data on inputs and participation, but also decision and management processes and data relating to and defining the curricular-instructional treatment defined units of participation.

C. Economic and social consequences--envisioned to be focused on a set of generic consequences which could be tracked to both cause and increased specificity according to the units of participation. For example, knowledge was expected to be a generic category, which would be more specifiable by treatment unit and order of consequence involved.

D. Selected data analyses--several specific data analyses were expected to emerge as having priority interest as consequence categories were matrixed against units of participation (e.g., race or income level of participants by unit of participation).

E. Recommendations--to address both general program concerns and the status and next steps of program evaluation.

F. Critique--an outside expert was contracted to critique the final report.

While this report would have met the minimum requirement of the mandate, within the established time, as fortune would have it, no other segment of the evaluation was prepared to submit a report on the March 1979 due date. Consequently, the study was extended. Four-H staff took advantage of this to plan two field studies to empirically assess the exploratory study's findings.

The first was derived from a review of participants attributed consequences. This study was planned to accomplish four tasks as follows:

1. Explore the nature and degree of participants' preference for 4-H.
2. Explore participants' perceptions of the 4-H units of participation and the experiential elements attributed to them.
3. Provide evidence of participants' agreement or disagreement with consequences attributed in the exploratory review.
4. Provide 4-H staff with a field tested instrument for assessing participant perception of consequences by unit of participation.

The second study was designed to empirically test the conceptual framework of consequences. This study would involve the consultation of an expert on evaluation instrument design to assist a number of state 4-H youth programs in the empirical measurement of consequences in each category. This would greatly

strengthen the evaluation findings by validating the conceptual categories overall, and by providing additional data on issues of priority concern. Further, it would help state 4-H units begin to take advantage of the evaluation projects work by contributing to instrument development and learning first-hand how to better approach issues of consequence.

One of the main advantages to be gained from even this small amount of empirical measurement was expected to be realized in enhancing the credibility and adequacy of the conceptual frame for consequences and its attendant treatment and causal theory work. In point of fact, the evaluation was expected to produce two products. A report on consequences in response to the Congressional mandate and the groundwork for a new type of national 4-H evaluation program.

When national scale experimental designs are impractical, and where decentralized program evaluators methods and variables differ so as not to be additive, the approach adopted here may contain the basic for an alternative solution. By building broadly conceptual program structure categories (admittedly containing some variance), it utilizes accepted and proven theories of educational cause and effect to predict consequences for verification. The next step would be for program staff to proceed from this beginning to develop a general causal model for the 4-H program. Backed by an adequate descriptive base, such a model could provide a general context within which innumerable specific variables and designs could be aggregated to build national reports which would provide externally acceptable explanation of cause. Local evaluators, because they could rely on a central causal model, would not need to worry over experimental designs unless they wished to, or had utilized a curricular-instructional treatment which did not fit any in the model aready.

In spite of the methodological boost given the originally intended design by the windfall opportunity to add new empirical data, we are convinced that the original design represents an adequate response to the charge.

ISSUES OF SCALE

It is interesting to observe that an evaluation design that does not involve collecting significant new empirical data is immediately challenged as a somehow inadequate evaluation. Yet, the originally planned report would have clearly met the requirement of the mandate given, within the less-than-twelve-month timeline. Of course such a design is rather unconventional for contemporary evaluation work and would be expected to be subject to some questions without subsequent field work to validate it. While no illusions are held about its level of scientific confidence, it is a reasonable and effective design for approaching such a broad and complicated charge (all social and economic consequences) at such a large scale within a broad-aimed education program, within limited time and cost constraints. If, as suggested earlier, such charges do become more prevalent, this exploratory method should prove of interest to others.

This design was selected in response to a particular set of evaluation problems, however, to the extent that larger scale evaluations tend to more often face similarly conflicted mandates and similar difficulties in defining and ordering an appropriate field of inquiry, we might expect designs for large scale evaluations to lean similarly to more general and exploratory methods. If so, then the early suggestion that different concerns, issues, and therefore evaluation designs might be related to scale of inquiry deserves further thought.

This experience in designing a national scale evaluation for 4-H suggests that there may be differences associated with large and small scale evaluations which can pose problems for both study designers and users. Some of the differences suggested include the:

A. Nature of the Evaluation Question: Due to the probable interests and identities of evaluation sponsors at each scale, large scale evaluations may tend toward assessment of underlying funding policy or rationales, while smaller scale studies may tend toward assessment of program operations. Since policy questions are frequently less appropriately narrowly definable, levels of preferred methodological rigor may suffer if such questions are to be addressed as intended.

B. Degree of Program Variance Involved: The more program units that must be accounted for, the greater the degree of program design and practice that must be accounted for in evaluation design and methods used. This complication is likely to be more pronounced in studies of nonformal education programs in which curricular-instructional treatments are less standardized.

C. Data Management and Collection: While it is clear that this task would be more costly and complicated as scale increases, the effect is compounded by increases in the degree of variance in program design and practice that must be accounted for.

D. Effect of Time and Resource Constraints: While larger, more complex evaluations are normally expected to take longer and cost more, it is very easy to underestimate how much more. For example, the sample survey that is so quick and inexpensive at the small scale can quickly become ungainly and cost in excess of $100,000 at the national scale. Cost and time constraints seem more likely to effect design and methodological choices in larger scale studies.

E. Stakeholders' Attitudes and Behavior: Major increases in evaluation scale seem to increase stakeholders' fears and suspicions, exacerbating any normal tensions. This is particularly likely as the degree of basic funding policy implication is perceived to increase. In

such an atmosphere, it seems likely to predict increased controversy over study design and results as scale increases. Another potential function of scale seems to involve local units commitment to the studies' needs and results. Larger scale efforts may lose much "volunteer" help as well.

While most of these differences do not appear to be individually necessary or inherent functions of scale of inquiry, to the extent that they correlate with scale, they will have profound implication for design and methodological choices, and appropriate expectations for study results. We think there is adequate reason for evaluators, evaluation theorists, and evaluation sponsors to reconsider methods and expectations for large scale evaluations, particularly post-hoc studies. It seems likely that as the scale of post-hoc inquiry increases, trade-offs between methodological rigor and adherence to questions of most interest will become more necessary and more potentially controversial.

PRAGMATISM AND INTEGRITY

It was this potential trade-off that was the essence of tensions expressed over integrity and pragmatism in the design of the national 4-H evaluation. Like the other concerns here associated with scale of inquiry, it seems likely that debate over this issue, in some form, will tend to intensify as scale of inquiry increases. Because both terms have strong connotations and are simultaneously interpretable to represent varied interests, evaluators confronted with them should be methodical and explicit about meanings. The experience of the national 4-H evaluation team was that there was never any disagreement over integrity or pragmatism in the abstract, only in context of design and methodology decisions did this debate become heated.

In this case, the issue of emphasis of integrity or pragmatism of design turned out not to be the issue as much as re-establishing a failing dialogue between partners in an educational program. The problem was not so much to choose between these demands, as it was to define them and the interests and fears they stood for.

At the same time, there are some implications, relative to integrity and pragmatism, suggested here for those interested in large scale evaluations. To be pragmatic means to be practical, to get busy with the task at hand, using plain, accepted tools. While most evaluators seek to be pragmatic, evaluators with large scale charges should have particular reason to be slower to make design decisions and begin work. As has been suggested, to define the task at hand from the perspective of most past experience is to project on the basis of smaller scale evaluation problems and designs, which may not be most appropriate to a large scale problem. To attempt to be too pragmatic too soon, may result in narrowing a study's scope to match familiar tools at hand before alternatives are fully considered. Some of the pressure to be pragmatic and "get to measuring" seems to originate in a popular perception of evaluation as simply the empirical measurement of some given parameters. Hence, a design such as the one originally proposed for the 4-H evaluation, which involved no collection of new empirical data on outcomes, can be quickly labeled "too academic" and not pragmatic. Yet, from another point of view, the original design was extremely pragmatic. It allowed initiation of useful inquiry immediately, utilized plain, unsophisticated tools, and was economical in terms of time and other resources. Designers of large scale evaluations should assume that their problem is not the same as most evaluation problems and avoid being urged into actions which assume too many design decisions too soon for the sake of pragmatism.

At least equally important are the implications of scale for design integrity. Integrity refers to the quality of honesty, of meeting moral expectations. This term can carry even stronger connotations of praise or criticism than does pragmatism. As was noted in the 4-H experience, concerns over design integrity can mask lobbying over whose perspective should prevail in study design. To the extent that larger scale evaluations tend to have more explicit policy and funding implication and involve a wider spectrum of program sponsors and stakeholders, this issue will be more problematic in some form.

One of the more problematic qualities of concerns for integrity, is that they can (and do) pull simultaneously in different directions. Policymakers tend to ask somewhat different questions of evaluators than would program practitioners, as exemplified in the charge to 4-H. One concern for design integrity is whether or not the study will give priority to addressing exactly the question of interest to the evaluation sponsors. Questions phrased, as 4-H's was, from a policy and fiscal perspective are not prone to easy definition or measurement. Furthermore, they tend not to be the questions of most interest to program designers and practitioners. The result is likely to be, as was the case here, that program staff will attempt to discredit the appropriateness and/or practical measurability of the sponsors question. In this event, evaluators should expect to be confronted with the view that integrity of design concerns the academic preferability of the study's methods. In effect, that it is better to redefine the question of interest so that more acceptable (i.e., more scientifically certain) measures can be brought to bear. The example posed in this paper is a case in point. To maximize scientific certainty, either the question of cause would have had to be foregone (probably via assumption) or else the range of consequences would have had to be drastically narrowed. Both alternatives were suggested in the name of design integrity.

Many current evaluation strategies argue for involving program stakeholders in some form of Delphi Technique interaction to get agreement on what should be measured. In fact, the 4-H evaluation team engaged in some activity to this end. However, mutual agreement was not readily forthcoming and in the final analysis it is evaluators themselves who must answer for the integrity of their efforts. Integrity in our view came down to attempting to answer the evaluation sponsors question of interest. A design was chosen which would provide for embracing the question of interest as fully as possible and the methodological trade-offs and necessary caveats about findings were explained to and accepted by the sponsor. There is no doubt about the threat implied in policy-based questions. Yet, for the most part, they are reasonable and appropriate requests for an accounting of return for investment. Program integrity can only suffer by avoiding such accountability. While program staffs' fears and concerns over design limitations, evaluation question and results interpretations, and levels of confidence of findings need to be openly and fairly addressed, there seems little room to doubt the priority that should be assigned to the evaluation interests and needs of those who support public programs and who also fund their own accountability studies. No evaluation question posed by a program sponsor (or participant) is really unreasonable or invaluable. From this perspective, any move to dilute or redefine an evaluation question once posed and clarified by an evaluation sponsor, should be considered a threat to design integrity. This does not cede to the evaluation sponsor the right to dictate design or methods. But, neither does concern for design integrity allow methodological preferences to infringe on the evaluation question at hand. As implied earlier, compromise and trade-offs characterize all evaluation design. But in larger scale, policy oriented studies, the issues are more volatile and often more clouded.

In this, as in questions of pragmatism, evaluators facing large scale evaluations will be well advised to proceed deliberately, without undue haste, and expect controversy. The likelihood is that, as with the national 4-H study, all concerns will not be satisfied in any first effort anyway. Compromises made between competing interests should have utility in assisting program staff and administrators pursue future evaluation in terms acceptable to and useful to program sponsors.

FOOTNOTES

1/The other four programs being: Agriculture and Natural Resources, Home Economics, Community and Rural Development, and Nutrition.

2/Nineteen hundred and seventy-seven farm bill, PL 95-113, Title XIV, Section 1459.

3/A good example is the National Headstart Program assessment. Westinghouse Learning Corporation, 1969, The Impact of Headstart: An Evaluation of the Effects of Headstart on Children's Cognitive and Affective Development, Bladensburg, Md., Westinghouse Learning Corporation.

4/Scriven, M., "Pros and Cons About Goal Free Evaluation," Evaluation Comment, Vol. 3, No. 4 (Dec. 1972).

5/"Issue Based Questions Relating to the Extension Evaluation," Evaluation of Economic and Social Consequences of Cooperative Extension, SEA/Extension, USDA, Washington, D. C., 1980.

6/MacDonald, J. and D. Clark, "Critical Value Questions and the Analysis of Objectives and Curricula," in Second Handbook of Research on Teaching, Travers (Ed.), Rand McNally, Chicago, 1973.