THE DESIGN, DEVELOPMENT, AND FIELDTEST OF AN EVALUATION FRAMEWORK FOR SHORT-TERM TRAINING PROGRAMS

By

Kent Jeffrey Sheets

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology, and Special Education

1983

ABSTRACT

THE DESIGN, DEVELOPMENT, AND FIELDTEST
OF AN EVALUATION FRAMEWORK
FOR SHORT-TERM TRAINING PROGRAMS

By

Kent Jeffrey Sheets

The study described in this dissertation originated from the need to identify an evaluation framework capable of assessing the impact of short-term training programs, specifically faculty development programs. An extensive review of the literature indicated that no appropriate evaluation approach existed. Although much of the literature on short-term faculty development programs reported that faculty development activities were successful and effective, results were based largely on self-reported and satisfaction data. This evidence was considered suspect by many authors. Therefore, a majority of the short-term training programs in existence were not evaluated in terms of impact on participants. This study was conducted to determine if an evaluation framework suitable for evaluating the impact of short-term training programs on participants could be developed and how well this framework would function when applied to an existing short-term training program.

An optimal evaluation framework for short-term training programs was designed eclectically by selecting elements and concepts from models and methods identified in the literature. The framework was fieldtested by evaluating a faculty development program involving 14 family physicians. Numerous methods were used to collect reaction, cognitive, and behavioral data from multiple information sources.

A metaevaluation was designed and conducted to assess the effectiveness of the fieldtest evaluation. An evaluator self-report, interview of the program directors, and analysis of evaluation procedures were used to gather data about the practicality, utility, and adequacy of the fieldtest procedures and outcomes.

The study's four conclusions are:

1. The Program had an impact on the participants and the framework documented the impact.
2. The most effective and efficient evaluation procedures were the End-of-Week Evaluations, final debriefing session, and videotape rating scale.
3. Discrepancies in evaluation results should be expected when qualitative and quantitative data are gathered from a variety of information sources using different evaluation procedures.
4. The evaluation framework is not useful for the purpose of providing immediate formative evaluation information to decision makers.

Recommendations for further research were presented and implications of the study for educational practice were discussed. In conclusion, a revised matrix of the evaluation framework was provided. The revised matrix reflected the results of the study.

To my wife, Barbara, for her support, understanding, and love throughout the writing of the dissertation.

ACKNOWLEDGEMENTS

A study of this nature is the product of the efforts and influence of individuals too numerous to recount. Family, friends, teachers, colleagues, and fellow students have all contributed significantly to this effort and I wish to acknowledge their support and contributions.

I am indebted to my parents for instilling in me at an early age a respect for knowledge, a zest for learning, and an appreciation of the importance of a job well done.

It was my great fortune to work with a committee composed of individuals who were my teachers, colleagues, and friends. Cass Gentry and Joe Levine were readily available for consultation and advice throughout the stages of my coursework and dissertation. Bill Anderson and Rebecca Henry supplied the idea and impetus for this study and provided tremendous support, encouragement, and guidance along the way. My chairman, Bruce Miles, has patiently nurtured me during my doctoral studies, keeping me on track, challenging me, and advising me.

As promised, I wish to thank Dan, Penny, Mike, and Eric for their time and effort in scoring the tests and videotape presentations. They contributed a substantial amount of their own valuable time to assist a fellow student and their efforts were truly appreciated.

Many thanks also to my friends and colleagues in the College of Osteopathic Medicine at Michigan State University and to my present colleagues in the Department of Family Practice at the University of Michigan. They were extremely supportive throughout the writing of the dissertation.

A special note of thanks is extended to my good friends from the pre-MSU era, Big Andy (How many pages?), and Larry and Sally (How are things in Ann Arbor, Ken?), among others too numerous to name.

A medal for service above and beyond the call of duty goes out to all those who typed various drafts of the dissertation, especially Steve and Marianne who labored long and hard at the CRT to type the final draft. Additional thanks go to Karen, Judy, Millie, and Blythe for their typing and editorial assistance along the way.

A final special word of thanks to the three people who made my years at MSU especially memorable and enjoyable. I feel fortunate to include Fred Benjamin and Eric Gauger among my friends, they helped make the low points tolerable, provided many high points themselves, and generally helped me maintain my "mental health."

Finally, thanks to the best friend from my stay at MSU, my wife, Barbara Beath. Barbara was also my editor, proof reader, and toughest critic. Most importantly, she was more than understanding when the dissertation had to come first. If and when she writes her own dissertation, I hope I can be half as understanding and supportive as she was of me.

TABLE OF CONTENTS

Tab

1
2
3
4

5
6
7
8
9
10
11
12
13

14
15
16

17
18
19
20
21
22
23
24

25
26
27
28
29

30

LIST OF TABLES

CHAPTER ONE

STATEMENT OF THE PROBLEM

INTRODUCTION

This dissertation reports the procedures, results, and conclusions of a study concerned with the identification of a validated evaluation framework which can be applied to short-term training programs. The study focuses on a potential solution to the growing problem related to the need to evaluate the impact of faculty development programs in post-secondary education.

A short-term training program is a program of from one hour to several weeks in length delivered to 50 participants or less. The program is designed to teach certain skills, techniques or content or to change specific attitudes or behavior. A short-term training program may be an independent program or it may be a component of a larger or longer program. Examples of short-term programs include workshops, seminars, intensive courses, orientation sessions, and conferences.

THE PROBLEM

Short-term training programs are conducted regularly throughout the United States and the rest of the world in a variety of institutions and organizations, including schools, corporations, hospitals, businesses, churches, and the military. Support for this statement is provided by

the existence of the large number of advertisements and notices for workshops, seminars, symposia, and other short-term training programs found in professional journals and periodicals.

In post-secondary education, short-term training programs are frequently used in faculty development programs directed toward the improvement of instruction and teaching. Gaff (1975) defined faculty development as "enhancing the talents, expanding the interests, improving the competence, and otherwise facilitating the professional and personal growth of faculty members, particularly in their roles as instructors" (p. 14). Other authors use the terms instructional improvement or teaching improvement to describe activities that fit Gaff's definition of faculty development.

Large amounts of time, effort, and resources have been and continue to be expended on the design and implementation of short-term training programs in a variety of settings and content areas. However, little is known about the impact of these programs because rarely are these programs systematically evaluated.

Forman (1980) attempted to explain why there is little or no history of systematic evaluation of training in business or industry.

> There appear to be three reasons which partly explain the low status of evaluation in training. The first is that, unlike education, a great deal of training occurs in the private, as opposed to the public sector. Since government and public foundations are not supporting these training programs, they cannot mandate evaluation.... Second, there is a general feeling (on the part of some people in business and industry) that educational methods often are not well suited to the real, everyday, outcome-oriented world of business. These people tend to distrust educational methods and techniques borrowed without adaption and revision; they want training evaluation to develop a character of its own.

> The third reason for the low use of evaluation in training is that the field of training is in a state of tremendous growth and development. Training is now a several billion dollar a year industry in the United States and growing at an incredible rate. It is interesting to note that when education was in a similar state of growth, evaluation was not very significant either. (p. 48)

Pratt (1979) reported findings similar to those reported by Forman and also commented on the lack of impact evaluations.

> The ultimate test of the quality of training is the impact the trained person has on some unknown future situation. This fact has been almost universally ignored in the evaluation of training.... Rather, evaluation of training has predominately focused on variables which deal with the actual process of training, including the instructor's style and technique, effectiveness of resources, and the student-instructor interaction. Additionally, evaluation typically focuses on the student's performance in relation to instructional objectives which, it is presumed, relate to knowledge and skill which will be useful at some future point in time. Often left unaddressed is the impact of training on practice and ultimately on the "system" in which the learner operates. (p. 350)

Forman and Pratt suggested that the status of training evaluation in business and industry needs to be improved substantially. They also noted that the focus of training evaluation efforts should shift from a heavy reliance on the assessment of participant satisfaction with the process of training to an assessment of trainee performance following completion of the training.

> Training is conducted to improve performance, and performance should be measured on the job, not just after the completion of a training program in the classroom setting. If the classroom benefits of training are not retained and transferred to the job, then training has failed to reach its full potential. Of all the range of evaluation activities, this stage is most important for the documentation of the effects and impact of training, and it is the stage which most clearly distinguishes educational from training evaluation. (Forman, 1980, p. 51)

Forman suggested that the systematic evaluation of training that is absent in business and industry is present in education. However,

according to a number of authors (Centra, 1976; Gaff, 1979; Hoyt & Howard, 1977; Levinson-Rose & Menges, 1981; Littlefield, Hendricson, Kleffner, & Burns, 1979; and Menges & Levinson-Rose, 1980) the literature of post-secondary education suffers from a shortage of reports of systematic evaluation of faculty development programs, including the short-term training programs often conducted within these faculty development programs. As Davis (1979) stated, "The major objective of all successful faculty development programs is to change the overt behavior of instructors in the classroom" (p. 125). However, the evidence supporting the successful change of overt teaching behavior of faculty development participants is for the most part based on satisfaction measures.

> One of the glaring problems in the evaluation of faculty development research is the tendency of authors to try and change teaching behavior but only evaluate the participants' reports of satisfaction with the course or their views of its relevance or usefulness. Almost invariably, courses are rated highly.... This does not tell the reader anything about what the participants have learned from the course. Even self reports of what the faculty believe they may have gained may be deceptive. (Stephens, 1981, p. 10)

Donnelly, Ware, Wolkon, and Naftulin (1972) suggested:

> Although there is great value to the satisfaction-type questionnaire, other kinds of data that permit the measurement of cognitive gain, attitudinal change, and ultimately behavioral change are crucial in evaluating any attempt at education. (p. 184)

Caldwell similarly reported in 1981:

> Not only is there a dearth of preservice training programs for teachers of adults, but also most existing training programs lack an evaluation component. The absence of evaluation procedures constitutes a serious deficiency in training program models. Far too often evaluation of training programs merely consists of questionnaires that elicit the responses of program participants. These questionnaires, or happiness indicators, measure the receptivity or responsiveness of the participants,

> but fail to measure the mastery of subject matter acquired by the participants or their attainment of program goals. (pp. 9-10)

In light of these reports, serious attempts should be undertaken to evaluate the impact of short-term training used within faculty development programs. These efforts should be designed to provide more rigorous data than the mere tabulation of participant opinions. Forman suggested the following guidelines for future evaluations of training:

> In training, evaluations will be more focused and less extensive. Training evaluations will have to be clearly linked to improving the program, documenting its effects, increasing its usefulness, or having some other demonstrable impact. Evaluation, in short, will have to be held more accountable for itself.
>
> Second, there will be changes in the data-gathering techniques used for the evaluation of training. In training, the emphasis must shift from survey techniques (questionnaires and interviews) and written tests to those that measure job performance, such as checklists, performance tests, observation scales, and role-play activities. Data must be gathered on what people can actually do, not just what they say they can do. (p. 50)

Similar guidelines could be formulated regarding evaluation of short-term training programs whether conducted in schools, churches, business and industry, or elsewhere. The problem is that no appropriate evaluation framework designed specifically for short-term training programs appears to exist.

Baron and Baron (1980) suggested a possible solution to the problem when they proposed that specific evaluation designs should be developed for different types of programs.

> We propose that evaluators abandon the aspiration to a single all-purpose research design. Instead, we suggest the development of several prototypes or ideal evaluation designs which fit different types of evaluation settings to varying degrees. As presently conceived, these prototypes could be generated both according to different conceptual orientations... and to the availability of time, money, and other resources. (p. 96)

Steele (1973) discussed the value of identifying an appropriate evaluation model and following an eclectic approach to its operationalization.

> In most instances you will select certain parts of a pattern for systematic evaluation. There's a growing push toward selective evaluation. For example, R.E. Brack of the University of Saskatchewan suggests that you take an eclectic approach--first identify the questions about the program that need to be answered and then select the parts of a particular model that can help deliver these answers without trying to systematically operationalize the complete model. In this situation, however, an understanding of the total pattern helps you keep the component that's receiving major attention within a total perspective of programming relationships. (p. 54)

Patton (1980) presented another viewpoint on this issue when he discussed comments made by a group of noted evaluators including Worthen, Stake, Stufflebeam, and Popham.

> The basic theme running through the comments of these evaluators was that their work is seldom guided by and directly built on specific evaluation models. Rather, each evaluation problem is approached as a problem to be solved--and the resulting design reflects their thinking about the problem as opposed to an attempt to carefully follow a prescriptive model. In effect, these experienced evaluators are describing how the practice of evaluation research requires more flexibility than is likely to be provided by any single model. (p. 58)

The differing viewpoints represented in the previous three quotations are indicative of the controversial nature of the issues related to the design and use of evaluation models. Cronbach, Ambron, Dornbusch, Hess, Hornik, Phillips, Walker, and Weiner (1980) contributed to the controversy and listed the following among their "Ninety-Five Theses":

> 55. Much that is written on evaluation recommends some one "scientifically rigorous" plan. Evaluations should, however, take many forms, and less rigorous approaches have value in many circumstances. (p. 7)

In view of the controversy noted above, the need to evaluate the impact of short-term training programs, and the absence of any evaluation

models designed specifically for short-term training, it is suggested that an evaluation framework for short-term training programs be designed. This evaluation framework should provide a mechanism that allows its users to conduct comprehensive evaluations of the outcomes of short-term training. At the same time, the evaluation framework should be flexible enough to be adapted to specific settings and allow its users to be eclectic in their operationalization of the framework.

PURPOSE OF THE STUDY

The study reported in this dissertation was conducted to determine whether an evaluation framework for short-term training programs could be developed and successfully implemented. As indicated earlier, short-term training is a popular training format. A great deal of time and resources have been and continue to be expended on short-term training programs with little or no assessment of their impact except for measures of participant satisfaction. It is becoming increasingly clear that individuals responsible for planning and implementing short-term training programs must also provide evidence that their programs are producing the desired impact on the ultimate target of the programs. It is assumed that an evaluation framework for short-term training programs would be of great interest to a number of these individuals.

LIMITATIONS OF THE STUDY

The evaluation framework resulting from this study was designed and developed based on a review of the literature of evaluation models and methodology. The framework was then applied to an existing short-term training program, a faculty development program for family practice

physicians. The evaluation conducted on the faculty development program served as the fieldtest of the evaluation framework. The evaluator shared the results of the fieldtest with the two program directors of the faculty development program by means of a written evaluation report.

There are several limitations to this study. The evaluation framework was fieldtested with one particular short-term training program. The purpose of the evaluation was to determine if this program had an impact on its participants. The framework was fieldtested on only one group of participants in one specific type of short-term training. There was no control group against which this treatment group was compared. Thus, the concepts of internal and external validity were of great importance when considering the limitations of this study.

Campbell and Stanley (1966) made a distinction between internal and external validity by defining internal validity as "the basic minimum without which any experiment is uninterpretable" (p. 5). In contrast, external validity was concerned with the question, "To what populations, settings, treatment variables, and measurement variables can this effect be generalized?" (p. 5). The internal validity of the fieldtest of the evaluation framework was addressed in this study by attempting to control the classes of variables that Campbell and Stanley identified as potential threats to internal validity. All possible precautions were taken throughout the process of developing and administering the evaluation instruments and while scoring and analyzing the data to minimize the effects of these variables. The external validity of the fieldtest results, or the validity of the inferences that could be made beyond the fieldtest, was partially established by the fact that the type of training evaluated in the fieldtest was a commonly used approach to faculty

development for physicians. While the fieldtest results may have limited generalizeability to other populations and programs, there was sufficient external validity to make inferences related to other faculty development programs for physicians.

The inferences concerning the effectiveness of the evaluation framework as a mechanism for measuring the impact of short-term training programs were more limited. The evaluation framework was not tested on other types of short-term programs or with programs with different content. Thus, inferences could be made only to the evaluation of similar programs with similar populations.

Based on this study it is difficult to claim that the specific short-term training program evaluated in the fieldtest would be similarly effective with a sample composed of non-physicians. Likewise it is difficult to propose that the evaluation framework would be similarly effective with a program with different content, length, or teaching strategies. However, stronger conclusions and recommendations can be made concerning whether or not the short-term program had an impact in this particular situation and whether or not the evaluation framework was effective when applied to this particular short-term training program.

RESEARCH QUESTIONS

The following research questions were formulated to direct the study:

1. What specific problems were encountered in the field-test of the evaluation framework?
2. Was the evaluation framework practical in its use of resources?
3. Was the evaluation framework useful in providing information to the decision makers?

4. Were the methods and instruments used during the fieldtest of the evaluation framework technically adequate?

5. Were the methods and instruments used during the fieldtest of the evaluation framework conducted in an ethical manner?

A metaevaluation, an evaluation of an evaluation, was designed and conducted by the evaluator to answer the research questions and assess the quality of the evaluation conducted during the fieldtest. In this manner the effectiveness of the evaluation framework was assessed as well. The evaluator, with the assistance of the program directors and established evaluation standards and criteria, evaluated the process, procedures, and results of the fieldtest.

DEFINITION OF TERMS

Behavioral data:

> information related to the performance of short-term training program participants in a simulated or on-the-job setting.

Cognitive data:

> information related to the knowledge and skills of short-term training program participants.

Evaluation:

> the determination of the impact of a program upon the program participants with the purpose of providing information to decision makers for planning, implementing, rejecting, and/or improving the program.

Evaluation framework:

> a set of conceptual components and guidelines to be utilized in the design, development, and implementation of evaluations.

Faculty development:

> enhancing the talents, expanding the interests, improving the competence, and otherwise facilitating the professional and personal growth of faculty members, particularly in their roles as instructors. (Gaff, 1975, p. 14)

Fieldtest:

> a step in the systematic development of a process or product in which the process or product is used in a setting that approximates the ultimate setting in which the process or product is to be used.

Impact:

> the effect of program participation on a participant and/or the participant's organization in terms of changes in the participant's cognitive knowledge, behavior, performance, and/or attitude.

Metaevaluation:

> the process of delineating, obtaining, and using descriptive and judgmental information about the practicality, ethics, and technical adequacy of an evaluation in order to guide the evaluation and publicly report its strengths and weaknesses. (Stufflebeam, 1981, p. 151)

Reaction data:

> information related to the satisfaction of short-term training program participants with the content, instructors, and activities of the program.

Short-term training program:

> a training program lasting from one to several weeks, composed of 50 participants or less that is designed to teach certain skills, techniques, or content or to change specific attitudes or behavior; may be an independent program or a component of a larger program.

Training:

> activities conducted with the purpose of helping participants (trainees) learn specific skills, techniques, methods, or attitudes to help improve their performance, usually in a job-related setting.

ORGANIZATION OF THE DISSERTATION

In Chapter One the problem was outlined and described. Research questions were presented and key terms were defined.

The review of related research in Chapter Two examines the literature on evaluation of faculty development programs in medical education and higher education, evaluation models, evaluation methodology, and metaevaluation. The material presented in this chapter serves as the source of information for the design phase of the study.

In Chapter Three the evaluation framework is presented with an explanation and rationale for the methods used to conduct the fieldtest of the framework. Procedures for evaluating the fieldtest, the metaevaluation, are outlined.

The results of the fieldtest are presented in Chapter Four. The metaevaluation results are also provided in this chapter.

In Chapter Five the dissertation is summarized and the results of the fieldtest and metaevaluation are discussed and interpreted. Conclusions are drawn and recommendations for further research are suggested.

CHAPTER TWO

REVIEW OF THE RELATED LITERATURE

INTRODUCTION

This review examines research literature on the evaluation of faculty development activities in medical education and higher education, evaluation models, evaluation methodology, and metaevaluation. Information presented in this chapter was used to design the evaluation central to the study. As a result, the chapter includes a discussion of the strengths and weaknesses of existing evaluation models and methods and metaevaluation models as they pertain to the usefulness of these models and methods for the evaluation of short-term training programs.

EVALUATION OF FACULTY DEVELOPMENT PROGRAMS IN MEDICAL EDUCATION

Faculty development programs have been popular in medical education for a number of years. Stephens (1981) conducted a review of the literature related to faculty development in medical education. Her review encompassed more than 40 articles and books dedicated to research on faculty development activities for medical teachers. An area of medical education that has used faculty development workshops in recent years is a new medical specialty area, family medicine or family practice.

> The establishment of family medicine as the newest medical specialty set the stage for the resurrection of the family doctor. The years from 1969 to present have witnessed an

> explosion of interest in this distinctive form of medical practice. (Canfield, 1976, p. 911)

This "explosion of interest" translated into medical school graduates selecting family medicine residency training and a resultant search for family medicine faculty by medical school administrators.

> Since no reservoir of experienced family physicians has existed to meet the demand for faculty during the past ten years, most faculty members in family practice training programs entered teaching after a period of 10 to 20 years in either group or solo practice. (Ramsey & Hitchcock, 1980, p. 421)

Faced with the problem of hiring faculty with little or no teaching experience, departments of family medicine have been forced to rely on faculty development programs to train teaching faculty. The workshop has become a technique frequently used in these programs.

> The search for an effective means to meet the faculty development needs of family medicine faculty revealed that workshops are a frequently used and effective method of promoting faculty development in general. (Bland, 1980, p. 8)

While there was little doubt about the accuracy of Bland's statement that workshops were a frequently used method of faculty development in family medicine, her comment concerning the effectiveness of workshops required further examination. Much of the research cited by Bland in support of that statement was based on self-reported data and satisfaction measures, rather than on objective outcome or impact measures. There was little evidence of evaluation of actual changes in participants' behavior due to short-term training in the literature of faculty development in medical education.

For example, an article by Bland, Reineke, Welch, and Shahady (1979) presented results of a study of the effectiveness of the two-to-three day workshop format for faculty development in family medicine.

> Clearly, the two-to-three day workshop format can result in enduring perceived change in faculty members' teaching, research, and administrative abilities. Also, participants report they have changed and/or increased their faculty activities at home as a result of the workshop. (p. 458)

Impact was not measured by observation of performance or by cognitive tests, but by self-report data collected from program participants who "perceived" and "reported" changes in ability and activities. As Stephens (1981) pointed out:

> Some type of systematic observation of teaching behavior is probably more useful than self report when assessing the impact of a workshop on teaching skills.
>
> This is not to suggest that ratings of faculty satisfaction with a workshop or course are not important. It is certainly crucial to please the consumers of a service. But this suggests measuring participants' satisfaction, using a rating scale of pre-post gains in teaching behavior as well as getting feedback on the structure of the workshop. (p. 10)

Evaluation efforts should go beyond the collection of satisfaction and self-report data if the impact of programs on participants and the participants' organizations is to be determined. Stephens addressed this issue and commented on the scarcity of such efforts.

> Generalization of change to outside the workshop is an important concern in evaluation. It is also one that has been widely neglected. A few authors (e.g. Bland, 1979) asked the participants to report how their behavior has changed. This method has all the problems that any self-report measure does. Irby et al. (1976) used a self-report measure, but strengthened it considerably by also observing the lectures of the participants at a later date. This is a practice that needs to be encouraged to establish the usefulness of faculty development.
>
> Once behaviors have generalized, they also need to be maintained. A change that lasts only for a few weeks or is exhibited only when a teacher is being observed is not a useful accomplishment. Follow-up contacts in faculty development research are as rare as attempts to assess generalization. (p. 11)

Other research cited by Bland as proof of the effectiveness of faculty development workshops was examined. Three studies cited by Bland

(Adams, Ham, Mawardi, Scali, & Weisman, 1974; Koen, 1976; and Wergin, Mason, & Munson, 1976) relied heavily or solely on self-report data or were not concerned with workshops as an instructional format. Only one reference, Donnelly et al. (1972), reported results based on the use of tests to measure attitude change and cognitive learning. A subsequent study by Bland and Froberg (1982) also reported positive results of faculty development workshops, but these results were based primarily on participant self-ratings.

> The primary data gathering instruments were the participant questionnaires (PQs), which asked for participants' self ratings of their abilities before and after the workshop or seminar. Because of their advantages in cost and efficiency, self-assessments are often seen by evaluators as the method of choice. Generally, self-assessments show moderate correlations with achievement or performance measures. It appears, however, that people may rate their own abilities somewhat higher than is warranted by their performance tests and also somewhat higher than they are rated by others, such as peers, superiors, or subordinates. (Bland & Froberg, 1982, p. 540)

Further examination of the literature of faculty development activities in medical education yielded mixed results. Joorabchi and Chawhan (1975) reported that by "using experiential learning methods in small groups with little or no didactic presentation, it was possible in a short time to change long-held educational views of diverse groups of medical educators" (p. 40). Pre- and post-tests of attitudes were used in this study to arrive at those results.

A study by Warburton, Frenkel, and Snope (1979) used evaluation approaches including interviews, videotapes, and self-assessment measures. Some positive impact was shown in reducing anxiety and increasing comfort among faculty participants in activities related to teaching family medicine. A study by Walls (1979) used objective tests to measure impact of a faculty development program on family medicine

faculty. Positive results were reported, but no observation of behavioral change or other impact measures were examined.

A study reported in 1980 by Lawson and Harvill used self-reports of attitude change along with ratings of videotaped teaching performances to evaluate the effectiveness of a faculty development program for residents. "The results of the study described here indicate that short training programs can produce significant, observable improvement in physicians' teaching behavior" (p. 1003). No mention was made in the report of any attempts to measure cognitive change in the participants.

A two-year-long faculty development program at the Michigan State University College of Osteopathic Medicine was evaluated with data gathered from program staff, faculty participants, and faculty non-participants. Although positive results were reported, the evaluation was based entirely on self-report data and there was no evidence presented that any observation of faculty using the content of the workshops was conducted. No mention was made of the use of cognitive tests for evaluation purposes (Bell, Hunt, Parkhurst, & Tinning, 1979).

Faculty development activities are well documented in the literature of medical education. Workshops were frequently used in these faculty development programs, especially those conducted with family medicine physicians. However, there was little or no evidence found in the literature that these faculty development activities were sufficiently evaluated in order to assess whether or not participants in these activities had actually changed their teaching behavior as a result of their participation.

EVALUATION OF FACULTY DEVELOPMENT PROGRAMS IN HIGHER EDUCATION

The next section of the review of the literature focuses on the evaluation of faculty development programs in higher education, particularly those that may be classified as short-term training programs. Levinson and Menges (1979) reviewed the research literature on improving college teaching and reported less than encouraging results. "The literature on teaching improvement in higher education is larger than we had expected when we began this review. It is also of lower quality than we had hoped" (p. VIII-1). Levinson and Menges examined six major categories of methods of fostering faculty development, but had some pertinent comments to make about workshops and seminars.

> Perhaps the most frequent but least carefully evaluated instructional improvement activities are workshops and seminars.... A number of courses to train graduate teaching assistants have been systematically evaluated. Activities for experienced faculty, on the other hand, are typically evaluated rather informally by questionnaires distributed at the close of an event or soon thereafter. Participants are likely to be asked how they felt about the activity and what they learned from it. These comments, at least as described in reports and published articles, are usually positive, but permit no conclusions about impacts which persist beyond the event itself. (p. IV-1)

In a subsequent work, Menges and Levinson-Rose (1980) stated again that "there have been virtually no adequate studies of the impact of workshops" (p. 2). In 1981, Levinson-Rose and Menges suggested the following guidelines for assessing impact.

> Because the most common data for evaluating workshops are participant satisfaction ratings (sometimes termed the "happiness index"), we note problems of such estimates. When studies assess satisfaction and skill at preworkshop, end of workshop, and delayed posttest, the happiness index is known to be seriously misleading.... From such research we extrapolate several guidelines for workshop assessment, guidelines seldom followed in research we reviewed: 1) both immediate and delayed tests should be made... and 2) if participants' self-assessments are to be accurate, they should refer to specific behaviors. (pp. 409-410)

Littlefield et al. (1979) supported the findings of Levinson and Menges and stated that "systematic evaluations of faculty development programs are difficult to find" (p. 4). Littlefield et al. also cited the following quotation from Hoyt and Howard (1977) to support that statement.

> In summary, the literature is extremely sparse and the studies reported are uncommonly simplistic. Apparently, participants in faculty development programs have generally expressed satisfaction with them, a finding of doubtful meaning. There is some evidence that teaching methods may change in directions considered desirable by teaching authorities. No dependable evidence regarding impact on students was reported. (Hoyt & Howard, 1977, p. 2)

Centra (1976) conducted a survey of colleges and universities in the United States to determine the status of faculty development practices in post-secondary education. A total of 756 institutions responded to Centra's survey. Of those, only 14% reported that they had evaluated their faculty development programs or activities, 33% had performed partial evaluations, while half of the programs had not been evaluated.

> A dozen or so respondents forwarded copies of their program evaluations. Judging from these, questionnaires or interviews with samples of faculty members were commonly used. Although such methods can prove helpful in tapping faculty reactions to particular services, or in ascertaining faculty awareness of a program, more sophisticated designs are probably needed to deal with such issues as accountability and the actual effects of various activities. (p. 42)

Gaff (1979) pointed out the dearth of information on the impact of faculty development programs. "While the literature of faculty development is replete with descriptions and analyses of programmes, little evidence has been gathered about the impact of these programmes on participants or on their institutions" (p. 242). Gaff went on to state that emphasis has been on the establishment of faculty development programs rather than on their evaluation. The evaluations that have been

conducted have been rather simplistic; participant reactions, annual reports, visits by outside evaluators, and case studies prepared by insiders or outsiders. These evaluations told more about the operation of the program that its outcome.

Gaff and Morstain (1978) suggested the possible problems that could result from relying on such happiness measures rather than observing faculty development participants in action following interventions. "It is one thing for faculty to give a generally positive assessment of their experiences, even indicating specific benefits of teaching improvement activities, but it is quite another for them to actually do something different in their teaching" (p. 78).

The literature provided little empirical evidence that faculty development programs in higher education have made an impact on participants. Most evaluation efforts appeared to stop when the activity was over and did not attempt to observe the participants' behavior in actual or practice application situations following the faculty development activities.

EVALUATION MODELS

An examination of the literature on evaluation and evaluation models indicated there were numerous definitions and models of evaluation in existence. Worthen and Sanders (1973) compared eight different models, each with a different definition, purpose, and key emphasis. Steele (1973) examined over 50 different evaluation models, approaches, and frameworks. Other authors (Borich & Jemelka, 1981; Britan, 1978; House, 1978, 1980; and Taylor, 1976) categorized existing models according to philosophy, purposes, assumptions, and other criteria. However, the

authors each had their own terminology and category scheme and rarely did they coincide or agree. Volumes have been written on the definition, purposes, and methods of evaluation, but there has been little consensus among the experts in the field concerning definitions or categories of evaluations. However, far from working against the prospective evaluator, this lack of consensus among the experts can be used to the practitioner's advantage. As mentioned in Chapter One, experienced evaluators reported they rarely followed a specific evaluation model. They were more likely to modify a model or models to suit a particular situation.

> In many situations, rather than extensively adapting a particular approach, you might be better off to construct your own, borrowing the parts of other approaches that are most useful and building patterns and processes that are appropriate to your needs.
>
> Don't search for the one way to do evaluation. Do search for the range of approaches that will best address your varied needs in program evaluation. (Steele, 1973, p. 55)

Patton (1980) went beyond Steele's suggestion of eclecticism to propose that it is the difference between the actual practice of evaluation and the ideal conceptualizations of evaluation that often leads to more meaningful and useful models. Patton also discussed some new options now available to evaluators.

> In essence, the options open to evaluators have expanded tremendously in recent years. There are more models to choose from for those who like to follow models; there are legitimate variations in, deviations from, and combinations of models; and there is the somewhat model-free approach of problem-solving evaluators who are active, reactive, and adaptive in the context of specific evaluation situations and information needs. Cutting across the evaluation model options are a full range of methods possibilities, the choice in any particular evaluation to be determined by the purpose of the evaluation, and the nature of the evaluation process. (pp. 58-59)

Based on the comments of Patton and Steele, it appeared the evaluator was free to examine a variety of evaluation models and then select the aspects of a model or models that best suited a particular situation. Following such a procedure, several models were examined to determine which had components suitable for the purpose of designing an evaluation framework to be used with short-term training programs. A number of evaluation frameworks, models, and approaches are briefly described, with emphasis on their strengths and weaknesses appropos to this study. The terms framework, model, and approach were used interchangeably in much of the literature reviewed and are used in the same manner throughout the dissertation.

Scriven, Stake, Stufflebeam, Tyler, Alkin, and Grotelueschen are authors of the models discussed in this chapter. Over 50 evaluation models were identified and examined, and the models of these six individuals were selected because of their relevance to the study and their prominence in the evaluation literature. Additionally, several categories of models are described. A table that summarizes the strengths and weaknesses of the models is provided later in this section.

Scriven (1967) wrote philosophically about evaluation and compared concepts of evaluation such as, goals versus roles, formative versus summative, and comparative versus non-comparative. Two concepts applicable to the problem of assessing impact of short-term training were intrinsic and pay-off evaluation. Intrinsic evaluation involved an assessment of the instruments or materials used in the program, while pay-off evaluation examined the effects of the materials or instruments on program participants. Both kinds of evaluation were relevant to determining program impact. Aside from these two concepts, Scriven's

philosophical discussion of evaluation did not lend itself to the evaluation of short-term training programs. According to Worthen and Sanders, there were serious methodological problems in Scriven's approach to evaluation. There was no methodology provided for assessing the validity of evaluative judgments and the approach contained several overlapping concepts. Except for the concepts of intrinsic and pay-off evaluation, this approach was not well suited to the purpose of this study.

Stake (1967) presented a much more descriptive and prescriptive model of evaluation than did Scriven. Stake's model was devoted to describing and judging educational programs using a formal inquiry process. One of the components of Stake's model provided for the assessment of program outcomes using a systematic approach that allowed for the use of relative and absolute judgments by the evaluator. However, Stake also called for the use of explicit standards, which may not always exist when determining changes in performance, attitudes, or behavior. Worthen and Sanders suggested that Stake provided inadequate data collection methods in his model and that some of the distinctions made between different cells of the model matrix were not clear and sometimes overlapped.

Stufflebeam's model of evaluation (1968) was a comprehensive approach to an evaluation of the context, input, process, and product of educational programs. The components of the model related to process and product evaluation were particularly applicable to an examination of program impact since these components focused on program activities and outcomes. Stufflebeam's view of evaluation included the concept that evaluation provided information to decision makers. This concept was

pertinent to the problem of how to determine the impact of short-term training programs so that programs may be planned, implemented, rejected, and/or improved. However, while the process and product components of the model had some utility, the context and imput components were not of similar value since they were more concerned with the planning of evaluations, thus negating the use of the model in its entirety.

Tyler's approach to evaluation (1942, 1949) was clearly based on behavioral objectives and the assessment of whether they were being achieved by the learners. While this was a good measure of program impact, additional information was required to indicate whether learners used the content of the program, changed their behavior, or were satisfied with the program. Tyler's approach was central to the task of evaluating the impact of short-term training programs, but his approach was not comprehensive enough because it failed to consider other important factors related to the impact of short-term training.

Alkin (1969, 1972, Alkin & Fitz-Gibbon, 1975) presented a holistic approach to evaluation that was decision- and system-oriented. Alkin suggested that the impact of the program on other systems be examined using documentation and outcome evaluation. This concept was similar to Stufflebeam's notions of process and product evaluations. The value of Alkin's model lay in its attention to other systems that interact with the program and its participants. However, Alkin did not clearly delineate methods to be used within this model and the systems approach may be very costly and complex to implement due to the time and resources it requires.

Grotelueschen (1980) presented a comprehensive approach to program evaluation. He described a classification scheme intended to specify

evaluation questions and clarify relationships among those questions. Grotelueschen's approach included consideration of three purposes of evaluation (to justify, to improve, and to plan), four evaluation elements (participants, instructors, topics, and contexts), and four program perspectives (goals, designs, implementation, and outcomes). Of particular value to the study were Grotelueschen's descriptions of the three purposes of evaluation, the elements, and the outcome perspective. Grotelueschen's whole model was more complex and comprehensive than required for the specific purpose of assessing program impact. However, the concepts of determining the purpose and elements of evaluation, the formulation of sample questions, and the focus on outcomes appeared to be of particular value to the study.

The remaining models were grouped under two categories of models rather than attributing them to a particular author. The first category reviewed was the transactional approach (Taylor, 1976) or the illuminative (Parlett & Hamilton, 1976) or contextual approach (Britan, 1978), depending upon which author was doing the categorizing or describing. The models in this category were primarily characterized by an intensive study of the whole program. Evaluation methods used with these models included observation, interviews, analysis of program documents, and other qualitative methods. The use of qualitative methods within the models in this category was applicable to impact evaluation, but the extensive use of observations and analysis of documents focused on implementation rather than impact and did not appear to be useful as a complete approach.

The clinical approach to evaluation (Glaser & Backer, 1972) was similar to the transactional category of evaluation. Glaser and Backer

advocated a holistic systems approach which utilized subjective measurement, consultation, and feedback among its program evaluation methods. The subjective measurement methods were applicable to the study, but the consultation and feedback methods were implementation-focused and there was a notable absence of any mention of the use of objective measures within this approach.

The strengths and weaknesses of the models described in this section are summarized in Table 1. While this review did not cover all the evaluation models identified during the search of the literature, it has mentioned those that were considered to be most applicable to the evaluation of short-term training programs. No single model was identified that was suited to the task of evaluating the impact of short-term training. However, concepts or elements that were relevant to this study were identified for possible inclusion in the evaluation framework to be designed.

Although they did not describe models of their own, several authors' views of evaluation and evaluation models were of interest and value. Steele (1973) suggested that evaluation should be conducted to judge and form conclusions and should be used as a management tool. She also said that program evaluation should be considered a generic term and that evaluators should look beyond objectives and results during program evaluations. This view was consistent with one of Scriven's concepts, goal-free evaluation, which suggested programs be evaluated without the evaluator's knowledge of stated program goals and objectives. Steele also suggested that unintended outcomes and results be sought and analyzed in evaluating a program.

TABLE 1

EVALUATION MODEL COMPARISONS

MODEL/AUTHOR	STRENGTH	WEAKNESS
Scriven	Intrinsic & pay-off evaluation	No method for assessing validity of evaluative judgments Overlapping concepts
Stake	Systematic approach Outcome evaluation component	Called for explicit standards Inadequate description of data collection methods
Stufflebeam	Decision-oriented approach Process, product components	Expensive to fully implement Context, input components designed for planning evaluations
Tyler	Focus on attainment of objectives	Too narrow a focus
Alkin	Decision- & systems-oriented Concerned with outcome, impact	No clearly delineated methods High cost of implementation
Grotelueschen	Concepts of evaluation purposes, elements, & questions	Too comprehensive & complex to use entire model Focused extensively on program design & implementation
Transactional	Holistic approach Use of qualitative methods	Focused more on implementation than impact
Clinical	Holistic, systems approach	Implementation-focused Lack of objective measures

Patton (1978, 1980) called for qualitative, utilization-focused evaluation based on an eclectic approach to the process of evaluation. Patton also proposed the use of a holistic, naturalistic approach to evaluation in order to provide information to decision makers. This was similar to the purpose of the models developed by Alkin and Stufflebeam.

Among all the models examined, none focused exclusively on impact evaluation, although several (Alkin, Grotelueschen, Stake, and Stufflebeam) considered outcomes as a major component of their models. Bryk (1978) explained some of the problems inherent in an impact study:

> First, for the program to be effective all subjects do not have to move in a particular direction, on all dimensions, for each unit of time. Second, even if we could measure short-term changes with perfect validity, without an understanding from a clinical perspective of the individual program that generated the numbers, we may not know what values to place on them. As a consequence, we may be unable to interpret the results of the impact study.... Clearly, then the questions we ask and the methods we employ must be carefully fitted to the nature of the program under study. (pp. 51-52)

Corbett (1979) reported on the absence of literature on impact evaluation. "Numerous evaluations of training design, methods, and techniques, as well as student learning in terms of educational objective, have been reported, but very few on impact" (p. 347). In calling for impact evaluations, Pratt (1979) stated that "in impact evaluation, we are examining not just the impact of training but the relative impact of competing and complementing forces that potentially influence the agency, system, or practice under consideration" (pp. 351-352).

Hunt (1978) suggested an approach to determining who and what to evaluate when determining impact, but stopped short of suggesting methods to use. Grenough and Dixon (1982) proposed a "systematic measurement process designed to demonstrate to management whether or not those

trained use their experience" (p. 40). They described this process of assessing the impact of training in terms of evaluating the "utilization of training."

Grenough and Dixon suggested that utilization of training may be measured either directly or retrospectively. In the retrospective mode, surveys, telephone interviews, and on-site interviews were used with existing descriptive and quantitative data. Although rather simplistic in its methodology and not yet fully developed, this approach seemed to have some potential value for this study.

For a variety of reasons, no single established evaluation model was well suited to the task of assessing program impact. Some models were too complex or costly to use. Others were too narrow in focus. However, several of the models contained components or presented concepts useful for impact evaluation. The concepts taken from models that were most useful for this study included intrinsic and pay-off evaluation, a decision-orientation, a systems-orientation, a holistic viewpoint, and a focus on utilization.

Intrinsic evaluation was a useful concept since it suggested the value of examining the materials used in a program as a means of determining what the outcomes should be. Pay-off evaluation was relevant because it was concerned with impact and outcomes of program materials and activities. The concept of decision-oriented evaluation was appropriate because the definition of evaluation for this study included as its purpose the provision of information to decision makers. A systems-orientation was essential because impact may often best be assessed by gathering information from individuals in systems other than the systems in which the trainee functioned. The importance of a

holistic viewpoint was based on the notion that all systems and components of a short-term training program should be considered as a whole in order not to isolate or neglect certain variables or factors that might have important significance when determining impact of the program. Finally, the focus on utilization was relevant for it suggested the need to gather documentation that the content of the program was being used in the work setting.

A number of evaluation models and approaches were presented in this section of the literature review. The relevance of these models and approaches to the task of evaluating the impact of short-term training programs was discussed and strengths and weaknesses of each were identified. Finally, those components and concepts most useful for this study were identified and discussed. These components and concepts are reflected in the design of the evaluation framework outlined in Chapter Three.

EVALUATION METHODOLOGY

A prevalent theme found in literature on evaluation methodology was that quantitative methods have dominated research and evaluation studies in the past amd may need to be supplemented on some occasions by qualitative methods. Several authors (Cronbach et al., 1980; Filstead, 1979; Glaser & Backer, 1972; Patton, 1980; and Reichardt & Cook, 1979) suggested that studies be designed combining the two approaches rather than relying solely on one approach or the other.

The obtrusive nature of quantitative research methods was a major reason that some authors have suggested there are situations when qualitative methods may prove to be more effective in conducting program

evaluations. Glaser and Backer stated that "program evaluations do not always lend themselves to rigorously quantitative approaches" (p. 54). Patton (1980) supported Glaser and Backer and added that "on many occasions--indeed for most evaluation problems--a variety of data collection techniques and design approaches will be used" (p. 18).

Cronbach et al. listed the following among their "Ninety-Five Theses."

> 54. It is better for an evaluative inquiry to launch a small fleet of studies than to put all its resources into a single approach.
>
> 56. Results of a program evaluation are so dependent on the setting that replication is only a figure of speech; the evaluator is essentially an historian.
>
> 59. The evaluator will be wise not to declare allegiance to either a quantitative-scientific-summative methodology or a qualitative-naturalistic-descriptive methodology.
>
> 60. External validity--that is, the validity of inferences that go beyond the data--is the crux; increasing internal validity by elegant design often reduces relevance. (p. 7)
>
> 95. Scientific quality is not the principal standard; an evaluation should aim to be comprehensible, correct, and complete, and credible to partisans on all sides. (p. 11)

Reichardt and Cook discussed the potential benefits of using qualitative and quantitative methods together. They stated that "two method-types can build upon each other to offer insights that neither one alone could provide" (p. 21). Filstead supported Reichardt and Cook and stated, "Qualitative methods are appropriate in their own right as evaluation-assessment procedures of a program's impact. Program evaluation can be strengthened when both approaches are integrated into an evaluation design" (p. 45).

By no means were qualitative methods presented as the sole approach to evaluation research. Rather, as with the selection of appropriate components from various evaluation models, one was urged to consider qualitative methods as yet another means of conducting evaluation research.

> Quantitative approaches clearly may be warranted in some cases; however to maximize the utility of the data gathered to those who authorize its collection, and avoid damage to an on-going program, it may be useful to consider viable alternatives or supplements to standard quantitative or experimental methods. (Glaser & Backer, p. 54)

Filstead added, "Qualitative methods provide a basis for understanding the substantive significance of the statistical associations that are found" (p. 45).

Several other authors supported the use of multiple methods to evaluate the impact of programs.

> A carefully designed strategy using mixed, multiple measures seems desirable. Although no single measure may be individually strong, several measures taken together can create a total picture that reliably captures the efficiency of an individual program. If a program is effective, then predictable patterns of outcome information ought to occur across multiple measures. (Bryk, 1978, p. 40)

Posavac and Carey (1980) "recommended that evaluators use multiple variables from a single source because the evaluation of a single variable to be the criterion of success will probably corrupt it" (p. 54). Patton (1980) added that "multiple sources of information are sought and multiple resources are used because no single source of information can be trusted to provide a comprehensive perspective on the program" (p. 157). Cronbach et al. added, "Multiple indicators of outcomes reinforce one another logically as well as statistically. This is true for measures of adequacy of program implementation as well as for measures of changes in client behavior" (p. 8).

An example of suggested multiple criteria to be used to evaluate training programs was found in the literature on training and development in business and industry. Kirkpatrick's four criteria for the evaluation of the effectiveness of training programs were cited throughout the literature (Brethower & Rummler, 1977; Goldstein, 1974; Kirkpatrick, 1967; Laird, 1978; Otto & Glaser, 1970; and Wexley & Latham, 1981). Wexley and Latham described the four criteria this way:

> 1. Reaction criteria measure how well the participants like the program including its content, the trainer, the methods used, and the surroundings in which the training took place.
>
> 2. Learning criteria assess the knowledge and skills that were absorbed by the trainee.
>
> 3. Behavioral criteria are concerned with the performance of the trainee in another environment, i.e., the on-the-job setting.
>
> 4. Result criteria assess the extent to which cost-related behavioral outcomes have been affected by the training. (pp. 78-79)

Brethower and Rummler listed four potential levels of evaluation which were clearly based on Kirkpatrick's criteria.

> 1. Do trainees like the training?
>
> 2. Do trainees learn from the training?
>
> 3. Do trainees use what they learn?
>
> 4. Does the organization benefit from the newly learned performance?

In summary, the literature on evaluation methodology suggested various options related to the selection of evaluation procedures. These options included recommendations regarding types of methods to use, the value of multiple measures and sources of information, and criteria that could be used to evaluate various aspects of a short-term training program.

METAEVALUATION

The literature on metaevaluation was rather limited since it was a relatively new concept. Scriven first introduced the term in 1969 and he and Stufflebeam have been among its leading proponents.

> Theoretically, meta-evaluation involves the methodological assessment of the role of evaluation; practically, it is concerned with the evaluation of specific performances. (Scriven, 1969, p. 36)

> Good evaluation requires that evaluation enterprises themselves be evaluated. Evaluations should be checked for problems such as bias, technical error, administrative difficulties, excessive costs and misuse. (Stufflebeam, 1981, p. 147)

Metaevaluation was essentially defined as the evaluation of evaluations, but the term has had different meanings for different authors. "There are as many potential conceptions of metaevaluation as there are of evaluation itself" (Stevenson, Longabaugh, & McNeill, 1979, p. 38). Some authors limited the focus of the concept of metaevaluation. Cook and Gruder (1978) used the term "to refer only to the evaluation of empirical summative evaluations--studies where the data are collected directly from program participants within a systematic design framework" (p. 6). Stufflebeam placed no such restrictions on the term in any of his writings (1974, 1978, 1981). He suggested that just as there were formative and summative evaluations, there should also be formative and summative metaevaluations. Stufflebeam placed no limitations on the type of evaluations that could be evaluated in a metaevaluation study.

A term often associated with and confused with metaevaluation was meta-analysis. Scriven (1980) defined meta-analysis as "a particular approach to synthesizing studies on a common topic, involving the calculation of a special parameter for each" (p. 83). Numerous studies

on a common topic were analyzed together to look for trends or significance across studies. This kind of analysis was not a component of this study. For the purposes of this study the concept of metaevaluation is based on metaevaluation as perceived and defined by Stufflebeam (1981) as:

> The process of delineating, obtaining, and using descriptive and judgmental information about the practicality, ethics, and technical adequacy of an evaluation in order to guide the evaluation and publicly to report its strengths and weaknesses. (p. 151)

The metaevaluation procedures described in Chapter Three and the results presented in Chapter Four are derived using Stufflebeam's definition for guidance.

Since metaevaluation was a relatively new concept, literature on the topic was scarce. In 1974, Stufflebeam reported that:

> The state of the art of meta-evaluation is limited in scope. Discussions of the logical structure of meta-evaluation have been cryptic and have appeared in only a few fugitive papers.... The writings on meta-evaluation have lacked detail concerning the mechanics of meta-evaluation.... Finally, there are virtually no published designs for conducting meta-evaluation work. Overall, the state of the art of meta-evaluation is primitive, and there is a need for both conceptual and technical development of the area. (p. 4)

Seven years later Smith made a remarkably similar statement.

> There has been relatively little work done to date in the area of meta-evaluation...with most efforts having been focused on the development of formal evaluation standards. The practice of meta-evaluation holds great potential, however, for illuminating the nature of evaluation practice, highlighting the difficulties of performing evaluations, and fostering a concern for excellence in evaluation service. (Smith, 1981, p. 263)

Smith also reported that:

> Evaluators have consequently had little practice in conducting meta-evaluations and the literature on the subject is sparse.... The number of actual meta-evaluations is still very small and I know of no comparative studies of meta-evaluation procedures. (p. 266)

Stevenson et al. reported an "absence of empirical literature on metaevaluation in the human services" (p. 45). These authors also reported that "the literature on metaevaluation...has focused largely on the methodological soundness of an evaluation as the criterion for its worth" (p. 44). Stevenson et al. noted that in many cases evaluators were interested in evaluating not only the means or methods of an evaluation, but also in examining its ends or outcomes or impacts on the organization or the rest of society. However, examples of these kinds of metaevaluations were not found in the literature.

While little work has been done with metaevaluations, authors have suggested guidelines and models for metaevaluations. These authors included Stufflebeam, Cook and Gruder, and Millman (1981).

Cook and Gruder presented seven models of metaevaluation based on time of the metaevaluation, status of the data, and the number of data sets involved. These models were best suited for use with large-scale evaluations such as city-wide, state-wide, or nation-wide evaluations of curricula, instructional innovations, or other large-scale programs.

Millman presented alternative methods for metaevaluation such as criticism techniques often used in the arts and music. Millman also provided a checklist which could be used to evaluate evaluation programs and/or products. This checklist was based on a similar checklist, the Key Evaluation Checklist (KEC), which was outlined by Scriven in 1980. Heading #18 on Scriven's KEC, Metaevaluation, suggested that the other 17 items on the checklist could be applied to the evaluation while planning, implementing, and evaluating an evaluation. Millman's checklist asked similar types of questions concerning preconditions, effects, and utility

of the program or product and of the evaluation that was conducted of the program or product.

Since 1974, Stufflebeam's concept of metaevaluation has become more refined and further developed. As mentioned earlier, Stufflebeam suggested there could be both formative metaevaluations to guide the evaluation and summative metaevaluations to publicly report the strengths and weaknesses of evaluations. Stufflebeam (1981) also stressed that "metaevaluations must be a communication as well as a technical, data-gathering process" (p. 151). He considered metaevaluation to be both a process and a product.

Stufflebeam also outlined four categories of evaluation standards that should be used to plan, conduct, and evaluate evaluations. These categories were:

1) utility standards
2) feasibility standards
3) propriety standards
4) accuracy standards

The Joint Committee on Standards for Educational Evaluation built upon Stufflebeam's four categories and published Standards for Evaluations of Educational Programs, Projects, and Materials in 1981. This work detailed 30 standards within the four categories and proposed that the standards be used in planning, conducting, and evaluating evaluations. Many of these standards were similar to items on the checklists devised by Millman and Scriven. Included with each standard were an overview, guidelines, pitfalls, caveats, an illustrative case, and an analysis of the case.

Baron and Baron (1980) discussed the history of ethics, standards, and guidelines for evaluations and expressed some strong opinions.

> Whereas we feel that basic ethical principles for evaluation should be universal and absolute, we believe that methodological standards should be particular and relative, for when we get to issues of methodology, we are dealing with decisions constrained both by situational realities about what is possible and by the state of the art in regard to new research design, theory and statistical approaches. (p. 89)

As reported earlier, there were few published accounts of metaevaluation studies. Giesen (1979) reported in her master's thesis the results of the evaluation of a particular evaluation model. She established six criteria based on Stufflebeam (1974) and evaluated an evaluation model based on those criteria. Her results showed that with some minor additions the model could be extremely useful and effective.

Kennedy (1982) applied the evaluation standards developed by the Joint Committee to a three-year faculty development, curriculum revision project. She reported how the standards were used in four phases of the evaluation project, designing the evaluation, collecting the information, analyzing the information, and reporting the evaluation. She identified those standards which were extremely useful as well as those which seemed to be of little or no value for the individual project phases.

In summary, the literature on metaevaluation was limited, both in reports on how to conduct a metaevaluation study and in reports on outcomes or results of such studies. Different authors have developed checklists and standards which can be used to plan, conduct, and evaluate evaluations. As these checklists and standards are used, more reports should be generated and added to the literature. This study incorporated a metaevaluation design based on some of the literature just described. This metaevaluation is described in Chapter Three and the results are presented in Chapter Four.

SUMMARY AND IMPLICATIONS FOR THE STUDY

The literature on evaluation of faculty development activities in medical education and higher education, evaluation models, evaluation methodology, and metaevaluation has been reviewed and discussed. This review revealed no single evaluation model properly suited to evaluate the impact of short-term training programs, thereby partially explaining the lack of such evaluation reports for faculty development activities in both medical education and higher education. Based on the review of the literature on evaluation models and methodology, some evaluation procedures and methods suitable for the evaluation of the impact of short-term training programs were identified. These methods and procedures were utilized to design an evaluation framework that is presented in Chapter Three. In addition to the evaluation framework, a plan for the metaevaluation of the fieldtest of the evaluation framework is also presented in Chapter Three.

CHAPTER THREE

PROCEDURES AND METHODS

INTRODUCTION

In this chapter, the procedures used in the design, development, and fieldtest of the evaluation framework for short-term training programs are presented. Based on the literature review in Chapter Two, an evaluation framework, which is referred to as an "optimal" framework, was designed. The optimal evaluation framework is described with its components and options as it is intended to be used. The program evaluated during the fieldtest of the evaluation framework is described, including a matrix of the evaluation framework as derived from the optimal framework and applied in this particular situation. The matrix also includes the evaluation questions asked during the fieldtest. The instruments and analysis procedures used during the fieldtest are also presented in this chapter.

The remainder of the chapter is devoted to a description of the metaevaluation of the fieldtest. The research questions originally stated in Chapter One are presented again and the procedures used to answer the research questions are outlined.

AN EVALUATION FRAMEWORK FOR SHORT-TERM TRAINING PROGRAMS

The evaluation framework described in this chapter was designed and developed to provide a mechanism for evaluating the impact of short-term training programs on program participants. As defined in Chapter One, the evaluation framework is a set of conceptual components and guidelines to be utilized in the design, development, and implementation of evaluations. The major purpose of evaluations conducted using the evaluation framework for short-term training programs is to provide the information to decision makers for planning, implementing, rejecting, and/or improving short-term training programs.

The evaluation framework was designed and developed based on information gathered during the review of the literature. Factors considered during the review and subsequent design included the importance of assessing the program impact, the focus on providing information to decision makers, and the need to provide users of the framework as much flexibility as possible. One assumption considered during the review and design was the probability that users of the evaluation framework would not have access to participants in their short-term training programs prior to the beginning of the program. Thus, data could be gathered only during and/or after the program.

The five major components of the evaluation framework are presented in Table 2. These components were taken from the literature reviewed in Chapter Two and are discussed in greater detail in subsequent sections of this chapter.

TABLE 2

COMPONENTS OF THE EVALUATION FRAMEWORK

COMPONENT	SOURCE	RATIONALE
1. Type of data gathered	Kirkpatrick (1967) Brethower & Rummler (1977) Wexley & Latham (1981)	Different types of data are required to conduct comprehensive evaluations.
2. Who or what is assessed	Hunt (1978)	The object of evaluation efforts must be identified to facilitate the process.
3. Source of data	Patton (1980) Cronbach et al. (1980)	Multiple sources of data help provide more comprehensive, reliable evaluation information.
4. Method of gathering data	Bryk (1978) Posavac & Carey (1980)	Different methods of collecting evaluation data should be used depending on the situation.
5. Evaluation questions	Grotelueschen (1980)	Sample evaluation questions facilitate the formulation of questions for specific settings & programs.

1. Type of data gathered

This component of the framework sets it apart from most of the existing evaluation approaches, models, and frameworks. Based closely on Kirkpatrick's four criteria for evaluation, the three types of data to be collected when using the framework are:

1) Reaction (satisfaction) data

2) Cognitive (learning) data

3) Behavioral (performance) data

Most of the research previously cited addressed only one, or at most two of the three types. Reaction data were the type most frequently collected from participants. Self-reports of cognitive or behavioral change were also used for evaluation purposes, but there was little

evidence of the use of objective measures of cognitive or behavioral change.

> It is important to recognize that favorable reaction to a program does not assure learning. All of us have attended meetings in which the conference leader or speaker used enthusiasm, showmanship, visual aids, and illustrations to make his presentation well accepted by the group. A careful analysis of the subject content would reveal that he said practically nothing of value--but he did it very well. (Kirkpatrick, 1967, p. 96)

> Participant self-reports are the most common method of measuring change in management training programs. Unfortunately, most program evaluators and researchers believe the self-report to be among the least accurate and least consistent forms of measuring participant change. (Mezoff, 1981, p. 10)

2. Who or what is assessed

The purpose of the evaluation framework is to assess the impact of a short-term training program on its participants. This component is concerned with identifying who or what is assessed in order to determine program impact.

The information gathered with the framework is ultimately concerned with assessment of the content, activities, and resources of the program for the purpose of making decisions and is only secondarily concerned with aptitude and achievement of the participants. However, it is often necessary to assess the participants in order to obtain accurate program evaluation data. In the framework, the participants are the object of cognitive and behavioral data collection and the program is the object of specific reaction data collection.

3. Source of data

There are numerous potential sources of data which provide information about the participants and the program. Ideally, all those individuals in a position to comment on changes in participants resulting from the short-term training program should be considered as data

sources. Minimally, the participants, program faculty, and supervisors of the participants should serve as sources of data about the participants and the program. Other possible sources of data include the participants' subordinates, peers, students, clients, family members, or others with whom the participants interact while applying the skills and techniques learned during the program. If the program content includes activities related to the creation of certain products or materials, it would also be possible to use examples of products or materials developed by the participants as sources of data. Cronbach et al. (1980) and Patton (1980) were among the authors who suggested the use of multiple sources of information when conducting evaluation studies.

4. Method of gathering data

The data-gathering methods used within the framework may be quantitative, qualitative, or both. The methods need not be the same for each type or source of data. Various methods may be used to collect data to answer a single evaluation question or one method may be used to answer more than one evaluation question. The work of several authors (Baron & Baron, 1980; Bryk, 1978; Cronbach et al., 1980; Patton, 1980; and Posavac & Carey, 1980) provided the impetus for including this component in the framework.

Evaluation methods that may be used to evaluate short-term training programs include, but are not limited to, the following:

1) Interviews
   - telephone
   - personal
   - group

2) Questionnaires
   - semantic differential questions
   - open-ended questions
   - checklists

3) Tests
   - multiple-choice, true-false questions
   - essay, short answer questions
   - oral
   - simulations

4) Direct observation (live, videotapes, films, audiotapes)
   - checklists
   - rating scales
   - narrative accounts
   - diary

5) Participant and staff self-reports
   - checklists
   - rating scales
   - narrative accounts
   - diary

5. Evaluation questions

General evaluation questions are presented within this component to help guide the use of the framework. The inclusion of this component was prompted by Grotelueschen's (1980) use of similar questions in his evaluation model. The questions presented in the matrix of the optimal framework in Table 3 are examples of the types of questions that might be of value to users of the framework. The specific questions used in the fieldtest are presented later in this chapter.

MATRIX OF THE OPTIMAL EVALUATION FRAMEWORK

The five components of the evaluation framework have been arranged in a matrix in an attempt to graphically represent the framework. Within the structure of the matrix, different individuals, activities, and elements were placed in appropriate cells. The following abbreviations are used for elements identified in the matrix.

P - Participant
F - Faculty (program)
S - Supervisor of participant
S-R - Self-report
STP - Short term training program
O - Others related to the participant (subordinates, peers, clients, students, family members)
Q - Questionnaire
I - Interview
VT - Videotape
DO - Direct observation

The elements displayed in the matrix in Table 3 depict an optimal configuration of the evaluation framework.

USE OF THE EVALUATION FRAMEWORK

Suggested guidelines for the use of the evaluation framework for short-term training programs are outlined in this section. Options and requirements to consider when operationalizing the framework are discussed.

One of the requirements of the evaluation framework is that all three types of data--reaction, cognitive, and behavioral--be collected. To assess program impact on participants, it is important to assess cognitive and behavioral change in addition to participant satisfaction. Kirkpatrick (1967) is a major proponent of using multiple evaluation criteria or levels to assess the outcomes of training.

Another requirement is that the program and participants are assessed to gather reaction data while participants alone are the object of cognitive and behavioral data-gathering activities. The ultimate objective of the evaluation is to evalute a program's impact, but to conduct that evaluation it is necessary to evaluate participants as well as the program.

TABLE 3

MATRIX OF THE OPTIMAL EVALUATION FRAMEWORK

TYPE OF DATA	ASSESS-MENT OF	SOURCE OF DATA	METHOD OF GATHERING DATA	EVALUATION QUESTIONS
REACTION (SATISFACTION)	STP	P	I,Q	How satisfied with the STP were the P's?
		F	I,Q,S-R	How satisfied with the STP were the F?
		S	I,Q	How satisfied with the STP were the S?
COGNITIVE (LEARNING)	P	P	TESTS	How much of the content of the STP did the P's learn and retain?
		P	I,Q,S-R	How well did the P's perceive their own learning and retention of the content of the STP?
BEHAVIORAL (PERFORMANCE)	P	P	DO,VT	How well did the P's apply the content of the STP in a real or simulated setting?
		P	I,Q,S-R	What did the P's report having applied on the job? What did they expect to apply in the future?
		P	I,Q,S-R	How did the P's perceive their ability to apply the content of the STP?
		P	I,Q,S-R	Did the P's perceive any change in their role or function in their organization?

TABLE 3 (Cont'd.)

TYPE OF DATA	ASSESS-MENT OF	SOURCE OF DATA	METHOD OF GATHERING DATA	EVALUATION QUESTIONS
		F	I,Q,S-R,DO	How did the F perceive the P's ability to apply the content of the STP?
		S,O	I,Q,S-R	Did the S and/or O perceive any change in the P's? How did the S and/or O perceive the P's ability to apply the content of the STP?

Many options are available to the user of the evaluation framework when selecting data sources, data-gathering methods, and evaluation questions. It is within these three components that the user has the most flexibility. Ideally, decisions related to operationalizing these three components will be made together. The source of data and the method of collecting the data depend upon the evaluation question being asked. The type of data being collected and who or what is being assessed also influences the selection of the source of data and method of data collection. It is important for the user to view the components of the evaluation framework as interacting systems when determining data sources, methods, and evaluation questions.

The questions in the matrix of the optimal evaluation framework in Table 3 were formulated to provide examples of the type of questions that might be asked during an evaluation of a short-term training program. Specific short-term training programs such as the program evaluated in the fieldtest require the formulation of specific questions relevant to the characteristics and conditions of the program being assessed. Previous decisions concerning who or what is assessed, who or what the source of the data is, and what data-gathering methods have been chosen must be considered when constructing these questions if the framework is to be implemented according to its guidelines.

PROGRAM, SUBJECTS, AND SETTING

The program evaluated in the fieldtest of the evaluation framework was a two-week session of the Family Medicine Faculty Development Program (Program) conducted by the Office of Medical Education Research and Development (OMERAD) at Michigan State University (MSU). The Program

began operation in July 1978 with the support of a grant from the Bureau of Health Manpower, Public Health Service. The Progam had two major objectives. One was to identify and train new physician teaching faculty for family medicine training programs. The second objective was to help current family medicine faculty develop and/or refine their teaching skills. One component of the Program was a three-month teaching fellowship. The fellowship was offered to allopathic (M.D.) and osteopathic (D.O.) physicians who had completed or were near completion of a family medicine residency program and to family medicine physicians with one year or less of academic teaching experience. The fellowship offering evaluated began in September 1981 as a two-week session held on the campus at MSU. The fellows returned to MSU in January, March, and May for additional one-week sessions. The session conducted from Tuesday, September 8, 1981 to Friday, September 18, 1981 was the object of the fieldtest of the evaluation framework.

The subject matter presented in this session focused on principles and techniques related to teaching and learning in medical schools and residency training programs. The content was presented in a variety of instructional formats including workshops, seminars, and simulations designed and developed by expert medical educators from MSU and OMERAD. The September 1981 session satisfied the characteristics of a short-term training program outlined in Chapter One. The session lasted two weeks, was completed by fewer than 50 participants, was designed to teach specific techniques and content related to teaching and learning, and was part of a larger program. A schedule of the session activities is provided in Appendix A. The topics presented during the two weeks included:

elements of group development
principles of learning and motivation
clinical teaching technique
curriculum development in family medicine
issues in family medicine
producing audiovisual materials
teaching psychomotor skills
presentation skills
role of clinical supervision
perspectives in learning (learning theory and decision analysis)
constructive feedback in clinical education
asking and answering student questions

Fifteen fellows started the program, but one dropped out during the September session, reducing the size of the group to 14. These 14 fellows comprised the population and provided the data for the fieldtest. This group was the largest to complete the program in the short history of the Program. Of the 14 fellows, four were D.O.'s, three of whom were first year faculty members. The fourth was a second year resident. Among the 10 M.D.'s, two were faculty members and the remaining eight were second or third year residents. Eight fellows were from the state of Michigan, two from New Jersey, and one each from Arkansas, North Carolina, Ohio, and Oklahoma. One fellow was a female, the remaining 13 were males. Each fellow received a stipend to defray the costs of participation in the program.

FIELDTEST OF THE EVALUATION FRAMEWORK

After examining the content and activities of the September 1981 session, the evaluator designed a plan for the evaluation of the session. The matrix of that plan appears in Table 4 and represents the activities that occurred during the fieldtest of the evaluation framework.

The instruments used during the fieldtest are described in the following section of the chapter. The information collected with these

TABLE 4

MATRIX OF THE OPTIMAL EVALUATION FRAMEWORK AS APPLIED TO THE PROGRAM

TYPE OF DATA	ASSESS-MENT OF	SOURCE OF DATA	METHOD OF GATHERING DATA	EVALUATION QUESTIONS
REACTION (SATISFACTION)	Program	Fellows	End-of-Week evaluations Interview Final debriefing	1. How satisfied were the participants in the September session with the content, instructors, and activities of that session?
	Program	Faculty	Interview	2. How satisfied were the program directors with the content, participants, and activities of the September session.
	Program	Supervisors	Interview	3. How satisfied with the Program were the supervisors of the participants?
COGNITIVE (LEARNING)	Fellows	Fellows	Tests	4. What was the participants' level of cognitive knowledge of the September session at the beginning of the session?
				5. Was there a significant change in the participants' level of cognitive knowledge at the end of the session?
				6. Was there a significant retention of the cognitive knowledge of the content of the session by the participants six months following the completion of the session?

TABLE 4 (Cont'd.)

TYPE OF DATA	ASSESS-MENT OF	SOURCE OF DATA	METHOD OF GATHERING DATA	EVALUATION QUESTIONS
	Fellows	Fellows	Interview	7. How much additional study of the content of the session was undertaken by the participants between the end of the session and the administration of the delayed posttest?
				8. How did the participants perceive their own expertise in each of the content areas of the September session both before and after the session?
BEHAVIORAL (PERFORMANCE)	Fellows	Fellows	Interview	9. What types of skills or techniques did the participants use following the completion of the September session?
				10. What types of skills or techniques did the participants expect to use in the next six months?
	Fellows	Supervisors	Interview	11. How did the supervisors perceive the participants' ability to apply the content of the session?
	Fellows	Fellows	Interview	12. How did the participants rate their own performance in a series of three completed simulations or presentations?

TABLE 4 (Cont'd.)

TYPE OF DATA	ASSESS-MENT OF	SOURCE OF DATA	METHOD OF GATHERING DATA	EVALUATION QUESTIONS
	Fellows	Fellows	Interview	13. How did the participants rate their expected performance in a repeat of the series of three simulations or presentations?
	Fellows	Fellows	Videotape rating	14. How well did the participants utilize the content of the session related specifically to presentation skills in two different presentations?
	Fellows	Faculty	Interview	15. How did the program directors perceive the participants' presentation skills?
	Fellows	Supervisors	Interview	16. Did the supervisors perceive any change in the participants' teaching due to the session?
	Fellows	Fellows	Interview	17. Did the participants perceive any change in their own role or function in their home institutions due to the session?
	Fellows	Supervisors	Interview	18. Did the supervisors perceive any change in the participants' role or function in their home institutions due to the session?

instruments was summarized in a written evaluation report prepared by the evaluator. The evaluation report was shared with the directors of the Program following completion of the fellowship in May 1982.

It was left to the discretion of the program directors whether the evaluation information was used for summative or formative purposes as part of their decison making. Since the formal evaluation report was not distributed to the program directors until after the completion of the 1981-82 offering of the fellowship, use of the data for formative purposes was not possible until planning began for the 1982-83 fellowship period.

INSTRUMENTS

Several different instruments were used during the fieldtest of the evaluation framework. Some of these instruments were already being used by the Progam staff. Other instruments were developed and used for the first time during the study. The five different types of instruments used in the fieldtest are briefly described. Samples of the instruments appear in the appendices.

1. End-of-Week Evaluation Forms

These instruments (Appendix B) were designed and administered by the program directors at the end of both the first and second weeks of the September session. The focus of these instruments was on participant satisfaction with logistics, instructors, facilities, curriculum, and other programmatic elements. Participants were also asked to provide self-assessments of their competence in relation to selected session topics.

2. Pretest, posttest, delayed posttest

A pretest (Appendix C), posttest, and delayed posttest were administered to the fellows in order to assess entry level cognitive knowledge, changes in cognitive knowledge over the course of the program, and retention of cognitive knowledge six months later. These tests were developed using test items submitted by session faculty. These items were edited by the evaluator and the program directors to standardize item format. The items were then arranged as a 40-item short answer, essay question test. Only the session segments directly concerned with skills, techniques, and theories related to teaching and learning were tested. The same test was administered as the pretest, posttest, and delayed posttest.

The pretest was administered during the first day of the September session in a classroom setting. The posttest was administered as a take-home examination to be completed by the final day of the session. The posttest was administered following the completion of all the segments of the sessions represented by items on the test. The delayed posttest was also administered as a take-home examination, but it was administered during the March session of the fellowship, six months following the completion of the September session. All 14 fellows completed the three tests.

All three sets of tests were scored independently by two raters who were trained by the evaluator. The raters were both graduate assistants in OMERAD who had previous experience as teachers and trainers. The evaluator trained them to score the tests using model answers provided by the faculty who submitted test items. The raters gained practice scoring responses on a pretest completed by an individual who was not one of the

14 fellows. The evaluator explained the rating scale, which is provided in Appendix D, to the raters and provided guidelines for scoring each of the 40 items on the tests. The fellows' responses were prepared so that the raters did not know whether questions were from the pretest, posttest, or delayed posttest. All responses were transcribed and typed so handwriting would not interact with rating.

The raters' scores showed sufficient inter-rater reliability (pre-test: .77; posttest: .83; and delayed posttest: .81) on all three tests that the scores for both raters could be averaged together to arrive at a single score for each participant on each test. This averaging of scores permitted further data analysis such as computation of mean scores and standard deviations. The Spearman-Brown split-half formula (Hull & Nie, 1981, p. 256) was used to determine the reliability of each rater's scores and also to determine inter-rater reliability.

Content validity, "the degree to which the sample of test items represents the content that the test is designed to measure" (Borg & Gall, 1979, p. 212), of the test was established by asking each of the session faculty to submit test items for his/her segment. An item analysis was conducted for each test item and discrimination and difficulty indices were calculated for each item on each test. Test results and further discussion of content validity and item analysis are presented in Chapter Four.

3. Videotape rating scale

As part of the Program, the participants were required to give several presentations to the program directors and their fellow participants. This activity provided the participants an opportunity to practice using some of the content of the September session. These

presentations were videotaped as part of the program and allowed the participants the opportunity to evaluate their own skills and techniques. To determine how well the participants applied the content of the September session, the evaluator developed a 16-item rating scale (Appendix E) to rate two presentations given by each of the participants. The first presentation took place in January and was completed by 13 of the fellows. One fellow was unable to complete the presentation due to a medical emergency. The second presentation was videotaped in May. All 14 fellows participated in this activity.

Two raters were trained by the evaluator in the use of the rating scale. The raters were both doctoral candidates and graduate assistants in OMERAD. The evaluator explained the 16-item rating scale to the raters and gave them the opportunity to rate portions of two videotaped presentations not included among the tapes to be evaluated as part of the study. The raters discussed their ratings of these practice tapes with each other and the evaluator in an attempt to standardize their ratings. The raters viewed all the tapes and independently rated them according to criteria established during the training session. Due to the unequal length of the tapes, the raters were asked to rate only the first ten minutes of each presentation.

The inter-rater reliability coefficients were not as high as those coefficients obtained from the written tests (January presentation: .49; May presentation: .59), but were still of sufficient reliability for this type of behavioral measure to be of value. Again the scores were averaged together to arrive at a single score for each fellow for each

presentation. The Spearman-Brown split-half formula was again used to determine reliability.

The content validity of the rating scale was established by using session materials to help develop the 16 items. Objectives and principles from the session workshops on presentation skills and principles of learning and motivation were converted into rating items since these two topics were directly related to the process of preparing and delivering presentations.

4. Interview protocols

To gather additional information about satisfaction, learning, application of session content, and other variables, a series of interview protocols (Appendix F) was developed by the evaluator. The participants were interviewed by telephone by the evaluator using a protocol developed with the program directors. The interview protocol was also piloted on a former Program fellow. The fellows answered questions in a variety of formats, including rating their expertise, indicating frequency of use of materials and content, and providing open-ended comments.

The supervisors of the fellowship participants were also telephone interviewed by the evaluator. A similar interview protocol which was developed with the program directors was used to interview the supervisors. Again a variety of question formats was employed.

Finally, the evaluator interviewed the two program directors simultaneously using another interview protocol. All questions in this interview were open-ended.

5. Final debriefing questionnaire

As part of the regular evaluation system of the entire program, a final debriefing was conducted during the May session by the program directors. In addition to the standard written questions included on the questionnaire (Appendix G) administered during this debriefing session, a question concerning the evaluation activities was inserted to ascertain the fellows' reactions to the various evaluation procedures related to the September session and its activities.

After the fellows completed the written final debriefing questionnaire, a discussion was conducted by the program directors. The discussion focused on the participants' final reactions to the entire fellowship program. During this discussion, the September session was reviewed, thus providing additional reaction data.

ANALYSIS PROCEDURES

Since the evaluation methods were both quantitative and qualitative in nature, the analysis procedures were also quantitative and qualitative. Quantitative data analysis was required for the test data, rating of the presentations, and quantitative items on the interview protocols and End-of-Week Evaluations. The rest of the data was qualitative, notably the open-ended comments made during interviews, End-of-Week Evaluations, and the final debriefing.

For those items which were quantitative, frequencies were determined and where appropriate, descriptive statistics were calculated. For the ratings of the two videotaped presentations, descriptive statistics were computed for the individual items, each rater, and the average of the two raters. Descriptive statistics were computed for the three tests and a

univariate analysis of variance (ANOVA) was conducted using the Statistical Package for the Social Sciences (SPSS). The ANOVA compared the results of the pretest with the results of the delayed posttest and also compared the results of the posttest with the results of the delayed posttest to determine if there was any significant difference among the scores for the fellows on the three different tests.

Qualitative analysis of the open-ended comments was conducted by reading and grouping similar comments for each question and instrument. Those groups of comments which were made two or more times were reported in the summary of results for that instrument.

METAEVALUATION OF THE FIELDTEST

A metaevaluation was designed and conducted to answer the research questions originally stated in Chapter One.

1. What specific problems were encountered in the fieldtest of the evaluation framework?
2. Was the evaluation framework practical in its use of resources?
3. Was the evaluation framework useful in providing information to the decision makers?
4. Were the methods and instruments used during the fieldtest technically adequate?
5. Were the methods and instruments used during the fieldtest conducted in an ethical manner?

The metaevaluation, an evaluation of an evaluation, examined the quality of the process and product of the evaluation conducted during the fieldtest. The procedures used in the metaevaluation were based on the review of the literature in Chapter Two. Three different metaevaluation activities were conducted to assess the effectivenees of the evaluation

framework as it was applied in the evaluation of the September 1981 session.

Following completion of the fieldtest and distribution of the evaluation report to the program directors, the evaluator prepared a self-report that addressed the problems encountered during the fieldtest. The evaluator outlined difficulties related to the development, administration, scoring, and analysis of evaluation instruments. The self-report provided an overview of the problems faced by the evaluator during the fieldtest and answered the first research question.

The second stage of the metaevaluation was approached in two steps. The procedures used during this stage collected information used to answer the four remaining research questions. The first step was to identify evaluation standards related to each of the research questions. The source of the standards was the 1981 publication, Standards for Evaluations of Educational Programs, Projects, and Materials. After one or more standards were identified for each research question, specific questions were formulated to address the concepts within each of the standards. The specific questions were then arranged as an open-ended questionnaire (Appendix H).

As the second step in this procedure, the evaluator interviewed the two directors of the Program by telephone using the questionnaire. The information gathered during this interview provided a substantial portion of the data required to answer the final four research questions. Additional questions not specifically related to the research questions, yet concerned with the effectiveness of the fieldtest, were also asked during the interview.

The final stage of the metaevaluation consisted of an analysis of the evaluation procedures used during the fieldtest. With the assistance of the program directors and another educator experienced in conducting and evaluating short-term training programs, the evaluator identified five factors related to the effectiveness of evaluation procedures. The five factors are outlined in Table 5.

TABLE 5

EVALUATION FACTORS

FACTOR	DESCRIPTION
1. Direct utility of information for decision making	Ease of applying information to decision making
2. Time efficiency	Time required for developing, administering, scoring, and analyzing procedures and data
3. Resource efficiency	Personnel, materials, and equipment required for developing, administering, scoring, and analyzing procedures and data
4. Credibility	Reliability and validity of procedures and data
5. Data manageability	Ease of data representation, summarizing, and analysis

Each of the fieldtest procedures was rated in relation to each of the five factors. A scale from one to three was used to rate each procedure, with one meaning low, two meaning medium, and three meaning high. First the evaluator rated each of the procedures in terms of the five factors. Ratings were based on the results of the fieldtest. Then the evaluator interviewed each program director independently by telephone. The program directors rated each procedure one factor at a time. The evaluator also asked the program directors to explain their rationale

for each rating. The procedures were rated according to their actual performance in the fieldtest, not on their potential value or optimal performance capabilities.

This procedure helped the evaluator identify those procedures which were most productive in the fieldtest, an important step in determining the effectiveness of the evaluation conducted during the fieldtest. Results of this and other metaevaluation procedures are provided in Chapter Four.

SUMMARY

Based on an extensive review of the literature, an evaluation framework for assessing the impact of short-term training programs on participants was designed. An existing short-term training program was used to fieldtest the evaluation framework and an evaluation was designed for this particular program using the evaluation framework as a guide. The evaluation was conducted using various methods and sources of data. After the results of the evaluation were distributed to the directors of the Program, a metaevaluation was conducted to assess the effectiveness of the evaluation framework itself. The results of both the evaluation and the metaevalution are presented in Chapter Four.

CHAPTER FOUR

RESULTS

INTRODUCTION

The results of the study are presented in this chapter. The results of the fieldtest of the evaluation framework are provided initially as the 18 evaluation questions asked during the fieldtest are paired with the results related to each question. Results are grouped together according to the three types of data gathered during the fieldtest, reaction, cognitive, and behavioral data. Finally, the overall results of the fieldtest are summarized.

The other information presented in this chapter is the result of the metaevaluation of the fieldtest. The results for each of the three metaevaluation procedures are outlined in relation to the study's five research questions and then the overall results of the metaevaluation are summarized. In closing, a summary of the chapter is provided.

RESULTS OF THE FIELDTEST OF THE EVALUATION FRAMEWORK

The results of the fieldtest described in Chapter Three are presented in this section. Information presented in this section is based on the evaluation report submitted to the directors of the Family Medicine Faculty Development Program (Program). A copy of the introduction to the evaluation report is provided in Appendix I. The 18

evaluation questions originally presented in the matrix of the evaluation framework in Table 4 in Chapter Three are restated in conjunction with the data used to answer the questions. The questions and results are arranged according to the three different types of data collected, reaction, cognitive, and behavioral data.

Reaction Data - Questions and Results

1. How satisfied were the participants in the September session of the Program with the content, instructors, and activities of that session?

Three different procedures were used to answer this evaluation question. The End-of-Week Evaluations, participant telephone interviews, and final debriefing session provided the data summarized in the following paragraphs.

As a group, the participants indicated moderate to high satisfaction with the content, instructors, and activities of the September 1981 session of the Program. Comments gathered via the three procedures were consistent in substance and frequency. Sample comments, both negative and positive in nature, are provided in addition to noteworthy quantitative data.

The End-of-Week Evaluations consisted of three parts. The first part addressed eight different program aspects such as amount of reading and comfort of the room. The participants assessed each aspect in terms of three responses:

1) Keep the same
2) Increase
3) Decrease

Specific suggestions were also solicited for each of the eight program aspects.

In Part II, the participants responded to a series of statements related to the usefulness and quality of individual topics or activities. A five-point scale was used to rate each statement.

SA - Strongly agree
A - Agree
U - Uncertain
D - Disagree
SD - Strongly disagree

The items on the third part of the evaluation form were open-ended. For example, participants were asked to identify presentations that were most helpful, not relevant, and that gave the participants the most difficulty.

Some notable information was gathered with the End-of-Week Evaluation administered at the end of the first week of the September 1981 session. Ten out of eighteen respondents indicated a desire to increase the comfort of the room. (The total number of respondents varies according to item and part of the evaluation form. The total number is often greater than 14 because individuals not enrolled in the fellowship attended a number of the workshops offered during September's session. These individuals also completed End-of-Week Evaluations.)

On Part II, 16 out of 18 respondents indicated uncertainty, disagreement, or strong disagreement with the statement, "The session on curriculum development in sports medicine was useful." Nine out of eighteen individuals indicated a similar range of reactions to the statement, "I have a better understanding of the similarities and differences between allopathic and osteopathic family medicine."

On Part III, 14 of 17 respondents picked the presentation on principles of learning and motivation as the most helpful presentation of the week. "The presentation on principles of learning and motivation was

the most useful. It had some concrete ideas I can carry with me." The presentation on curriculum development in sports medicine was considered "not relevant" by 11 of 17 individuals. "The session on curriculum development in sports medicine could have been more relevant. It did not answer many of my questions on the subject."

The second End-of-Week Evaluation was administered at the end of the September 1981 session. The items in Part I were the same as for the first week. The statements in Part II referred only to the topics and activities of the second week. The questions in Part III included consideration of the entire two weeks and the upcoming January session.

At the end of the second week, only one person indicated a desire to increase the comfort of the room. The other 12 respondents suggested room comfort be kept the same. In addition, 7 of 13 individuals suggested the need to increase the amount of practice. Specific suggestions related to the increase in practice included, "Increase time, but try not to decrease information given out" and "Maybe more relevant practice examples from our own experience."

For nine of the ten statements in Part II, ten or more participants agreed or strongly agreed with the statements. Seven individuals agreed, three were uncertain, one disagreed, and two strongly disagreed with the statement, "I am more aware of my own thinking as a physician as a result of the discussion on perspectives in learning."

The presentations on teaching psychomotor skills, 6 of 13, and presentation skills, 5 of 13, received the most support as the most helpful presentation. "Those presentations involving our role playing and practice teaching were equally most helpful." Six of thirteen participants indicated the presentation on perspectives in learning was

not relevant. "I'm still not sure what the purpose of the perspectives in learning session was. It didn't seem particularly meaningful."

When asked what was missing from the session, three fellows suggested that changes be made in the content of the session segments related to clinical teaching and supervision. Content related to teaching residents and conducting hospital teaching rounds was requested.

When asked to provide their overall reactions to the two weeks, 13 of 13 fellows provided extremely positive comments. "Very positive. I feel I have learned a lot of techniques that I will utilize."

General impressions gleaned from the data indicated that the End-of-Week Evaluations were successful in eliciting positive and negative comments pertaining to content, positive comments on individual instructors, and constructive criticism related to session activities. Although not specifically articulated in this review of the results, comments related to these issues are distributed throughout the results provided in Appendix J.

Several items on the telephone interview questionnaire addressed the issue of fellow satisfaction with the September 1981 session of the Program (Appendix K). All 14 fellows indicated they felt the session helped them become better teachers. Seven individuals said the session gave them a conceptual framework to use when teaching. Five others responded that the session provided them information and skills they lacked before, while three fellows were pleased because the session gave them opportunities to practice presentation skills.

Each of the participants said they would recommend the session to a friend or acquaintance interested in becoming a teacher or a better

teacher. Three of the fellows had already recommended the session to someone.

Several of the fellows made comments during the interviews that were similar to comments made on the End-of-Week Evaluations. Three of the fellows stated that they still did not understand where the presentation on perspectives in learning fit into the session. "There wasn't anything taught except some theoretical things that I didn't find applicable at all to my situation."

Two participants indicated that the concrete, practical aspects of the session were most helpful to them. "I guess the things that were most helpful were the things that were concrete, practical, problem-oriented type things that could hang on the conceptual framework that was presented."

Dissatisfaction with the content related to clinical supervision and teaching was voiced again. "I think that the program would be improved if they could get a lot more reference points for how to teach residents rather than how to teach medical students."

The final procedure used to measure participant satisfaction was the final debriefing session conducted in May. Written responses were collected in the first stage of the debriefing before the program directors led a discussion reviewing the entire fellowship. Numerous comments related to the September session were made during both segments of the final debriefing (Appendix L).

When asked to provide an overall evaluation of the program, each of the participants expressed satisfaction with the program. "I feel that the program was valuable for me as a future family medicine educator and that I learned a great deal of relevant information that will be useful

to me in the future." A majority of the fellows said they would recommend the program to other physicians. "Overall, the program was excellent. I would recommend it to any family practice resident, regardless of whether or not the person was pursuing a faculty position in family practice. Teaching skills are applicable to many environments."

Two individuals reiterated the absence of important information relevant to clinical teaching and supervision. "Much of my time is spent on teaching rounds at the hospital. This was totally ignored in the program and needs to be addressed."

One of the questions on the written portion of the final debriefing was concerned with the participants' reactions to the evaluation procedures. The general consensus was that the End-of-Week Evaluations and telephone interviews were acceptable and valuable. The feedback on the cognitive tests was much less favorable. One fellow's comment typified the feelings of the group. "The pre/posttests were, in my opinion, worthless. The telephone interviews probably gave a better idea about what was being used.... The end of week evaluations were probably the best of the three. A long-term (one-year? five-year?) follow-up should be done."

Several comments pertaining to the September session were made during the discussion segment of the final debriefing. Suggested changes included dropping the presentations on issues in family medicine and perspectives in learning. The fellows expressed displeasure with the presenters of the curriculum development in sports medicine presentation, but stressed the idea that the topic of curriculum development should be kept in the curriculum. The other major suggestion was that more time be

provided for practice using skills such as giving presentations and producing audiovisual materials.

A tremendous amount of information was collected to answer this evaluation question. Much of the information was redundant and served to support comments collected using different procedures. The overall indications were that the participants were satisfied with the content of the session with the exception of three topics, perspectives in learning, issues in family medicine, and curriculum development in sports medicine. The fellows also suggested that new content be added to the sessions on clinical teaching and supervision. The segments that received the most positive comments were the presentations on principles of learning and motivation, presentation skills, teaching psychomotor skills, and producing audiovisual materials. Comments about instructors were favorable, except for the suggestion that the instructors for the curriculum development presentation be replaced. The activities of the session received positive reactions, especially those that provided opportunities for practice. The cognitive tests were not favorably received, but the participants responded favorably to the End-of-Week Evaluations and telephone interviews. It was noteworthy that comments provided six months (interview data) and nine months (final debriefing data) following the session were similar in substance and frequency to comments made at the end of the first and second weeks of the two-week session.

2. How satisfied were the program directors with the content, participants, and activities of the September session?

One procedure was used to gather information to answer this question. The evaluator interviewed the program directors simultaneously using a 12-item open-ended questionnaire. The interview was conducted following the completion of the May session of the Program. Comments

pertaining to the program directors' satisfaction with the content, fellows, and activities are provided among the interview summary presented in Appendix M.

The program directors agreed with the fellows' assessment of the weak areas of the curriculum. "I knew that perspectives in learning was a high risk going in." The problems with the presentations on issues in family medicine and curriculum development were also recognized by the program directors. They were also interested in the fellows' assessment of the content related to clinical teaching and supervision. "We expanded the clinical teaching component and one of my concerns was whether or not those three sessions held together as a unit." There was some dissatisfaction expressed concerning the clinical faculty who participated in the session, especially since two of the problem presentations were presented by clinical faculty.

The program directors were generally satisfied with the participants as a group. They characterized the group as being extremely oriented toward the practical aspects of the program, a change from previous groups. The directors were disappointed in the performance of three individuals who could not demonstrate the ability to apply the information presented during the fellowship. Despite their disappointment with the three individuals, the program directors were still pleased with the group as a whole. "I think we had a higher quality fellow this year, didn't take them as far, but the overall quality this year is higher than last year."

Activities were conducted to the satisfaction of the program directors. "The major project presentations were the most well thought out

statements they had made.... I think the presentations and major projects were clearly better than last year or any other year."

The program directors shared many of the same opinions as the fellows. The directors were disappointed with the performance of several individuals, but overall they were pleased with the content, fellows, and outcome of the session.

3. How satisfied with the program were the supervisors of the participants?

The supervisors of the participants were interviewed by telephone by the evaluator. Several of the questions asked during the interview pertained to the supervisors' satisfaction with the program. Composite results are provided in Appendix N.

The supervisors reported that 13 of the 14 fellows benefited their organizations in some way due to the fellows' participation in the Program. These benefits included the supervisors' reports that 12 of the fellows had shared information learned during the September 1981 session with them or other members of their organization. "I think it really benefited us because it exposed him to a lot of the concepts and ideas that come to bear in discussions that we have around problems and issues in the faculty meetings."

Eleven of eleven supervisors (three of the supervisors had two representatives of their organizations participating in the Program) said they would send another resident or faculty member to the fellowship in the future if the proper arrangements could be made. Two respondents viewed the fellowship as an opportunity for young faculty to grow. Two other supervisors were using the Program to try to accelerate the development of a young academic program. "I would like to have had this kind of experience myself at the time I left practice in 1974, because I

didn't know what I was supposed to know. I'm not sure I still know what I'm supposed to know, but I know I didn't know it then."

A general comment expressed by two supervisors was that the logistics of the program were very workable. However, two other individuals indicated it would have been helpful if the supervisors knew more about what they could do to facilitate the learning of the fellows.

In summary, the supervisors expressed their satisfaction with the Program. Their organizations benefited from the fellows' participation and the supervisors said they would encourage other individuals from their organizations to participate in the Program in the future.

Reaction Data - Summary

The reaction data gathered from the participants and program directors identified similar areas of satisfaction and dissatisfaction. The overall reactions of both the participants and the program directors were positive with the exception of the areas previously discussed. The supervisors were not as intimately involved with or aware of the finer details of the Program, but their reactions were positive as well.

Cognitive Data - Questions and Results

4. What was the participants' level of cognitive knowledge of the content of the September session at the beginning of that session?

A cognitive test consisting of 40 items was administered to collect data in response to this question. The results for the test and its seven subscales are presented in Table 6 and Table 7. The seven subscales represent the nine topics represented on the test. The items for the three presentations concerned with clinical supervision and teaching, clinical teaching technique, role of clinical supervision, and constructive feedback, were considered as one subscale, referred to as clinical teaching.

5. **Was there a significant change in the participants' level of cognitive knowledge of the content at the end of the session?**

The same test administered as the pretest was used as a posttest to gather data to answer this question. Posttest results are displayed in Table 6 and Table 7.

6. **Was there significant retention of the cognitive knowledge of the content of the session by the participants six months following the completion of the session?**

The cognitive test was administered to the group of fellows six months following the completion of the September session to answer this question. The test results are provided in Table 6 and Table 7.

TABLE 6

COGNITIVE TEST RESULTS

	PRETEST	POSTTEST1	POSTTEST2
Mean*	44.86 (37.88%)	68.32 (56.93%)	60.18 (50.14%)
Std. Dev.	8.34	11.29	10.54
Range	35.50 (23.50-59.00)	37.00 (46.50-83.50)	41.00 (34.00-75.00)

* 120 points possible

TABLE 7

COGNITIVE TEST SUBSCALE RESULTS

SUBSCALE	PRETEST	POSTTEST1	POSTTEST2
Elements of group development (12 pts.)	4.68 (38.98%)	4.54 (37.79%)	4.64 (38.68%)
Principles of learning & motivation (15 pts.)	5.50 (36.66%)	7.50 (50.00%)	6.93 (46.18%)
Clinical teaching (27 pts.)	10.93 (40.47%)	14.71 (54.49%)	13.96 (51.71%)
Audiovisual (12 pts.)	3.71 (26.45%)	7.86 (65.47%)	6.86 (57.14%)
Teaching psychomotor skills (21 pts.)	6.67 (31.76%)	13.64 (64.96%)	10.89 (51.86%)
Presentation skills (15 pts.)	8.36 (55.71%)	11.43 (76.18%)	10.29 (68.56%)
Perspectives in learning (18 pts.)	4.50 (25.00%)	7.96 (44.24%)	5.96 (33.13%)

Analysis of Cognitive Test Results

Following the scoring of the responses on the three tests, a univariate analysis of variance (ANOVA) was conducted comparing the results of the pretest with those of the delayed posttest. This analysis was done to determine if there was a significant difference between the pretest and the posttest. Another ANOVA was conducted comparing the posttest results to delayed posttest results to determine if the difference in the results of the two posttests was significant.

The ANOVA results are presented in Table 8 and Table 9. The alpha level for both comparisons was set at .01. The critical F value (1, 13 degrees of freedom) for the test was 9.07. Since the results of both comparisons were statistically significant, the difference between the

test results for the pretest and posttest was also statistically significant.

TABLE 8

ANOVA: PRETEST VS. DELAYED POSTTEST

Source	df	MS	F
Group	1	3286.44	
Error	13	74.75	43.96

TABLE 9

ANOVA: POSTTEST VS. DELAYED POSTTEST

Source	df	MS	F
Group	1	928.29	
Error	13	86.59	10.72

There are three assumptions that researchers try to satisfy when using ANOVA procedures. In brief the assumptions of the ANOVA are:

1) Independence of observations between and within groups

2) Normal populations

3) Equality of variance of the observations

In an evaluation such as the fieldtest of the evaluation framework, it is not always possible to meet all three assumptions. The small sample size and high probability of dependence among variables offer ample reason to criticize the robustness of these results. However, the differences between the means for the three tests do appear to be meaningful and of considerable value for the type of evaluation conducted during the fieldtest.

The mean results for the subscales of the tests appeared to follow the same trend as the overall test means, signifying an increase in

knowledge and retention of a meaningful proportion of that knowledge six months later. No statistical analysis was conducted with the subscale results. The three top subscale scores were presentation skills, audio-visual, and teaching psychomotor skills. The lowest two subscale results were elements of group development and perspectives in learning. The elements of group development subscale was the only subscale that did not follow the same trend as the overall tests. The scores for this subscale went down from pretest to posttest. Although the mean score increased from the posttest to the delayed posttest, the mean score for the delayed posttest was lower than the mean score for the pretest.

No criteria were established to determine levels of performance on the tests. Individuals wrere ranked according to their scores on the tests, but no judgments were made concerning the relative meaning of scores of 50 out of 120 and 65 out of 120, other than that one was higher than the other. The test data were presented to the program directors so they could make their own inferences about the performance of the fellows on the cognitive tests.

7. How much additional study of the content of the session was undertaken by the participants between the end of the session and the administration of the delayed posttest?

Two questions from the telephone interview with the participants addressed this question. The participants were asked if they had pursued additional study in any of the ten areas of the session related specifically to teaching and learning. One of the areas, asking and answering student questions, was not included among the topics on the cognitive test due to scheduling constraints that dictated its absence on the test. The topics, issues in family medicine and curriculum development in sports medicine, were not included on the test or the interview because

these topics did not incorporate principles and techniques related to teaching and learning.

The participants were also asked to identify which handouts or notes they had referred to since the end of the session six months earlier. Those fellows who indicated they had used their handouts or notes were also asked to estimate the frequency of use. Results are summarized in Table 10.

The only notable evidence of additional study was indicated by the information that 6 of 14 fellows did additional study in clinical teaching technique and the role of clinical supervision and 4 fellows did some additional study in presentation skills. The use of handouts was more evident. At least half the group reported using the handouts for five of the topics. The frequency of use for these five topics was also highest for all topics. The topic that received the least attention since the end of the September session was perspectives in learning. A discussion of these results as they relate to other results is presented in the summary of the cognitive data.

8. How did the participants perceive their own expertise in each of the content areas of the September session both before and after the session?

During the telephone interviews in March, each of the participants was asked to rate his or her own level of expertise in ten topics of the sessions. They were asked to first rate their expertise prior to the September session. Then they were asked to rate their expertise in the topic at the time of the interview. A scale from one to five was used, with one representing low expertise and five representing high expertise. The results are summarized in Table 11.

TABLE 10

ADDITIONAL STUDY AND HANDOUT USE

	ADDITIONAL STUDY		HANDOUT USE					
TOPIC	YES	NO	YES	NO	ONCE	2-3	4-5	5+
Elements of group development	2	12	5	9	4	1		
Clinical teaching technique	6	8	9	5	2	5	2	
Role of clinical supervision	6	8	8	6	3	5		
Constructive feedback in clinical education	2	12	7	7	3	3	1	
Principles of learning and motivation	2	12	6	8	3		3	
Teaching psychomotor skills	1	13	6	8		6		
Producing audiovisual materials	2	12	9	5	3	3	3	
Presentation skills	4	10	9	5		5	2	2
Asking and answering student questions	1	13	5	9	3	2		
Perspectives in learning	1	13	2	12	2			

TABLE 11

MEAN SELF-RATINGS OF EXPERTISE*

TOPIC	SEPTEMBER	MARCH
Elements of group development	1.88	3.28
Clinical teaching technique	2.67	3.82
Role of clinical supervision	2.30	3.82
Constructive feedback in clinical education	2.70	3.64
Principles of learning and motivation	2.50	3.21
Teaching psychomotor skills	2.44	3.71
Producing audiovisual materials	1.83	3.18
Presentation skills	2.25	3.79
Asking and answering student questions	2.45	3.39
Perspectives in learning	1.63	2.11

*5 points possible

The highest rated topics prior to September were clinical teaching technique, principles of learning and motivation, and constructive feedback. The lowest rated topics prior to September were perspectives in learning, producing audiovisual materials, and elements of group development.

When they were interviewed in March the participants were asked to indicate their current expertise in the topics. The top rated topics were presentation skills, clinical teaching technique, clinical

supervision, teaching psychomotor skills, and constructive feedback. The lowest rated area in March was perspectives in learning.

Statistical analysis was not conducted on these results, but the results are meaningful when compared to other results presented in this chapter. A discussion of these results is presented in the following summary of the cognitive data collected during the fieldtest.

Cognitive Data - Summary

An analysis of the cognitive data gathered during the fieldtest revealed some notable findings. Although not statistically measured, there was an apparent relationship between the test subscales, self-reports of expertise, and self-reports of handout use. There was also an apparent relationship between these cognitive data and a portion of the participant reaction data.

For example, the fellows scored best on the presentation skills on all three tests. Presentation skills was also the topic that ranked third in both ratings of expertise and handout use. It was also one of the most favorably received topics according to the satisfaction measures. Similar results across cognitive measures were found for the topics, teaching psychomotor skills, clinical teaching, and producing audiovisual materials.

On the opposite end of the spectrum, the session presentation on perspectives in learning fared poorly on all measures. That topic received the lowest ratings of expertise and handout use and the participants performed worst on that subscale on the pretest and delayed posttest, and second worst on the posttest. The topic also received more unfavorable reactions than any other.

In summary, there was high agreement across the cognitive measures at the top and bottom of the rankings. There also appeared to be a strong relationship between the cognitive data and reaction data for those topics at either end of the spectrum.

The cognitive data also indicated that the fellows performed better on the posttest than on the pretest. Although delayed posttest results were lower than posttest results, the participants performed better on the delayed posttest than on the pretest. Thus, there was a change in the participants' knowledge and the change was sustained over a period of six months.

Behavioral Data - Questions and Results

9. What types of skills and techniques did the participants use following the completion of the September session?

The participants were asked during the telephone interview to describe what specific knowledge or skills they were able to use in the six months since September. All 14 participants reported they were able to use some of the knowledge or skills learned during the September session. "I'm really utilizing a combination of almost all of them in my teaching in the clinic and the section on group development in my involvement in committees and other groups." Those topics which were identified at least five times are presented in Table 12.

TABLE 12

KNOWLEDGE OR SKILLS USED SINCE SEPTEMBER

TOPIC/SKILL	FREQUENCY OF RESPONSE
Presentation skills	11
Audiovisual production	9
Clinical teaching	9
Teaching psychomotor skills	7
Clinical supervision	6
Constructive feedback	6
Group development	5

10. What types of techniques did the participants expect to use in the next six months?

Another question from the telephone interview questionnaire served to gather the data to answer this evaluation question. All 14 fellows expected to be able to use content from the September session in the next six months. "I'm going to have a quarter time teaching position starting July and I'll be precepting then." Two of the participants reported their present situation did not provide many teaching opportunities, but they were expecting to be doing more teaching in the future. "I hope that in the next six months that I'll be in a position of doing more teaching.... My present situation just doesn't have very much opportunity to use a lot of these things." Topics which were identified at least five times are presented in Table 13.

TABLE 13

KNOWLEDGE OR SKILLS TO BE USED IN THE NEXT SIX MONTHS

TOPIC/SKILL	FREQUENCY OF RESPONSE
Clinical teaching	8
Presentation skills	8
Group development	7
Teaching psychomotor skills	7
Audiovisual production	7
Clinical supervision	5

11. How did the supervisors perceive the participants' ability to apply the content of the session?

A question from the supervisor interview asked the supervisors to describe any new knowledge or skills the participants had used since September. The supervisors' responses indicated that all 14 fellows used some new knowledge or skills. The use of clinical teaching skills and presentation skills were reported by five supervisors. "A couple of weeks ago I heard him present a talk. I think it was fairly evident from just watching him that he had picked up some skills in presentation of lectures." Four supervisors noticed fellows using skills related to group discussions.

12. How did the participants rate their own performance in a series of three completed simulations or presentations?

During the interview the fellows were asked to rate their overall performance in three completed activities. A scale from one to five was used with one serving as the low end of the scale and five signifying the high end. The results of the ratings are provided in Table 14. The participants rated themselves highest in the clinical teaching simulation

and proposed research presentation. The lowest rating was for the practice teaching exercise.

13. How did the participants rate their expected performance in a repeat of the series of three simulations or presentations?

After the fellows assessed their overall performance in each of the three completed activities, they were asked to assess their expected performance if they were to repeat the experience. The same five-point scale was used and the results are presented in Table 14 with the results of the original ratings.

TABLE 14

MEAN SELF-RATINGS OF PERFORMANCE*

ACTIVITY	COMPLETED	EXPECTED
Clinical teaching simulation	3.41	4.11
Practice teaching exercise	3.23	4.19
Proposed research presentation	3.45	3.88

*5 points possible

Although not statistically analyzed, the results showed positive change for each of the three activities. It was interesting to note that the least change was predicted for the presentation activity, even though it was rated highest initially. Comments made among the reaction data indicate a desire by some fellows for more practice giving presentations. Perhaps this desire was reflected in the ratings.

14. How well did the participants utilize the content of the session related specifically to presentation skills in two different presentations?

To answer this question, videotapes of two presentations given by the fellows were rated by trained observers using a 16-item videotape rating scale. The first videotaped presentation (VIDEO1) was completed

by 13 fellows in January 1982. The second videotaped presentation (VIDEO2) was completed by all 14 fellows in May 1982. Results are presented in Table 15.

TABLE 15

MEAN SCORES ON VIDEOTAPE RATING SCALE*

	VIDEO1	VIDEO2
Mean	40.35 (50.43%)	41.71 (52.14%)
Range	30.00 (21.00-51.00)	27.00 (25.00-52.00)

*80 points possible

The results indicated there was little difference between the mean scores for the two presentations. No statistical analysis was conducted since the presentations were conducted four months and eight months following the completion of the session. Participants were again ranked according to their performance on both of the presentations, but no criteria for determining appropriate or desired levels of performance were set. The data were shared with the program directors for use in their decision making.

Individual items on the rating scale were analyzed to identify specific areas of strength and weakness. The participants fared poorly on both presentations on items calling for interaction and consideration of the audience during the presentation.

> Related the presentation to the audience members' past, present, or future.
>
> Solicited ideas or questions from members of the audience.

The fellows also performed poorly on these items on one of the presentations.

Summarized main points or ideas of the presentation.

Varied the rate and pace of the presentation.

Responded to ideas or questions from members of the audience.

The participants' performance was rated highly for the following three items on both presentations.

Delivered, rather than read, the presentation.

Presented information in an organized, logical manner.

Showed interest in the topic and enthusiasm in presenting it.

Performance on the following item was rated highly on VIDEO2.

Spoke clearly and audibly.

These results demonstrated that the fellows needed additional work on making topics meaningful to the audience, varying rate of presentation, and involving and responding to the audience during the presentation. The fellows' presentation strengths were in organization, delivery and enthusiasm.

15. How did the program directors perceive the participants' presentation skills?

During the interview with the program directors, the presentation skills of the fellows were discussed. The directors observed both positive and negative changes in the presentation skills of individual fellows. "Some individuals used presentation principles we talked about while others that watched them didn't pick that up or chose not to do likewise." Three of the fellows performed at a substantially lower level than the rest of the group according to the program directors. Despite the poor performance of certain individuals, the program directors still concluded that this group was superior in the area of presentation skills to fellows of previous years. "When you look at all of them, they're

still a bit better than past groups. A couple are clearly low, but there's a couple of those in each group."

Since the presentations were given in September, January, and May, the program directors were able to observe change over time. "It was easy for me to see changes from September to January for four of the fellows." One program director indicated that 50% of the group had improved from September to May while the other director was more optimistic and suggested that as much as 70% of the group showed improvement. "I saw a lot of them make attempts to use overheads and some organization that I had not seen before."

16. Did the supervisors perceive any change in the participants' teaching behavior due to the session?

Two questions on the interview addressed this evaluation question. The supervisors reported they had observed 11 of the fellows teaching since September. The frequency of observations per supervisor ranged from once to five times or more. The types of teaching most frequently observed were clinical teaching, presentations, and group discussion.

The supervisors also indicated they were able to judge from these observations whether the participants' teaching behavior had changed. Five supervisors noted that the fellows were more comfortable and/or confident in their teaching. "He's more comfortable teaching, especially lecturing." Three supervisors noticed greater organization in the teaching of the fellows. "He seemed to be more organized. I think he was conscientiously and consciously using some approach and technique, particularly with facilitating discussions." Two of the supervisors attributed the changes in teaching behavior to the Program while two others felt that maturity and the passing of time may have been partially responsible for the change.

17. Did the participants perceive any change in their own role or function in their home institutions due to the session?

The participants were asked during the telephone interview if their role or function in their organization had changed due to their participation in the September session. Seven of the fellows reported a change of one sort or another. "I think the biggest change has been in my self-image as a teacher. I didn't give any credence to thinking of myself as a teacher, but after the September session I felt like it was legitimate to think of myself as a teacher and developing as a teacher." Four fellows reported they were trying to do more teaching or had changed their teaching style. "I've tried to do more teaching and have changed in that regard. So I think the way that I'm looked at by the faculty in my residency is slightly different." Two fellows said they were not involved in teaching at the moment. "I'm not in a position where I'm doing much teaching."

18. Did the supervisors perceive any change in the participants' role or function in their home institutions due to the session?

The supervisors provided information during the interview that responded to this evaluation question. The supervisors reported that ten fellows had changed their role or function since September. Two fellows were serving as coordinators of preceptor programs. "He's more involved. He coordinates my preceptor program and my community health and psychiatry clinical rotations. He has taken a much more active part in the clinical teaching." Supervisor comments also indicated that two fellows were more involved in research since September. "He has become more active in working with residents and in surveying patients as a beginning of a more active research activity."

Behavioral Data - Summary

The various methods used to collect behavioral data indicated that the participants, as a group, were using a number of the skills and techniques presented during the September 1981 session of the Program. There were occasions when data collected from one group were corroborated by data gathered from a different group or with another procedure. For example, the supervisors reported that the fellows were better organized in their teaching. Organization of presentations was one of the characteristics of the fellows' presentations as identified by the videotape ratings.

The information provided by the participants related to the knowledge and skills they used in the six months was supported by the reports of the supervisors. The participants reported using knowledge or skills related to presentation skills, clinical teaching, and group development among the seven topic areas they identified. The supervisors commented that they observed the participants using skills related to these three topics. The program directors also noted the participants' use of some presentation skills during their presentations given during the fellowship. Thus, there was agreement among sources of information for this particular aspect of the evaluation.

The participants all indicated they had improved as teachers due to the session. The supervisors were able to make a similar judgment about 11 of the 14 fellows. The program directors concluded that a majority of the fellows improved their presentation skills over the course of the fellowship, but the directors were extremely disappointed with the performance of three individuals. The three individuals also ranked near the bottom of the group on the videotape rating of the two presentations.

The discrepancy of views on the relative improvement of the three fellows is discussed in Chapter Five.

In summary, positive results were identified via the behavioral data collection procedures. There were occasions when results based on one data source or procedure were confirmed by another data source's or method's results. There was one occasion when there was disagreement across measures and procedures. In the following summary of the results of the fieldtest, the results for all three types of data collected are compared and major findings are presented.

SUMMARY OF RESULTS OF FIELDTEST

A vast amount of information was collected to answer the 18 evaluation questions of the evaluation conducted during the fieldtest of the evaluation framework. The important aspects of the results have already been presented within the categories of reaction, cognitive, and behavioral data. The major findings of the fieldtest are presented in this summary. The relationship of the results for the three types of data gathered are also considered.

The strengths and weaknesses of the content, instructors, and activities of the September 1981 session of the Program were identified and agreed upon in most instances by the participants and program directors. No major disagreements were evident in the reactions of these two groups. The supervisors also provided favorable reactions to the Program although by necessity their comments were limited in scope. Moderate to high satisfaction was reported by the 14 fellows, 2 program directors, and 11 supervisors.

Among the results of the cognitive measures, there was an apparent relationship between participant test performance, self-reports of expertise, and self-reports of handout use. There was also evidence that suggested this relationship extended to the degree of satisfaction with the topics as well.

The cognitive data also demonstrated that there was a meaningful change in the cognitive knowledge of the participants. This change occurred over the course of the two weeks and was still present six months following the session.

The behavioral data indicated the participants were using a number of skills and techniques related to the September session. Support for this finding was provided by data originating from all three sources of information and from the different data-gathering methods.

However, there was disagreement across sources as to the number of fellows who made noticeable improvement during the Program. The objective assessments of the program directors and the videotape rating results did not coincide with the subjective assessment of the fellows and the objective views of the supervisors. This discrepancy is discussed further in Chapter Five.

Across the three types of data there were some notable trends that deserve identification. The practical, concrete, skill-oriented presentations were most highly enjoyed. The content of these presentations was referred to the most, used the most, and used and learned the most successfully (according to cognitive test subscales and self-reports). The opposite was seen for the session on perspectives in learning which was viewed as theoretical and irrelevant.

In most instances, the information provided by the participants on different measures administered at various times over the nine-month period was consistent to the point of redundancy. There was sufficient reason to believe the data provided by the participants on the reaction measures and self-reports were reliable. The validity of the self-reports is discussed in Chapter Five. The data provided by the participants, program directors, and supervisors also were in agreement in most instances.

There was an apparent relationship between the performance of the participants on the three cognitive tests and the videotape ratings. Those individuals ranked at the top of the test results were likely to be highly ranked for the videotaped presentations. Similar results were found for those individuals who scored poorly on the cognitive tests. Scores and ranks for the fellows are presented in Table 16.

The results presented to this point have demonstrated that the September 1981 session of the Program was successful. The session participants learned new cognitive content, utilized important presentation skills in videotaped presentations, and were able to use knowledge and skills learned during the session in their teaching activities in their home institutions. The September session appeared to have impact on their cognitive knowledge and behavior.

RESULTS OF METAEVALUATION OF THE FIELDTEST

After the results of the fieldtest of the evaluation framework were collected and reported to the program directors, the metaevaluation of the fieldtest was conducted to collect data to answer the research questions stated in Chapter One. The summarized results of the

TABLE 16

COMPOSITE SCORES AND RANKS ON TESTS AND VIDEOTAPES

FELLOW	PRETEST*	RANK	POSTTEST1*	RANK	POSTTEST2*	RANK	VIDEO1**	RANK	VIDEO2**	RANK
1	23.5	14	56.5	11	34.0	14	21.0	13	40.0	10
2	41.0	10	46.5	14	46.0	13	41.0	8	52.0	1
3	48.0	5	69.5	8	61.5	7	48.5	2	41.5	8
4	39.0	13	73.0	6	75.0	1	41.5	6	34.5	12
5	52.0	2	81.5	2	73.0	2	41.5	6	52.0	1
6	39.5	12	79.5	3	63.0	6	32.5	12	25.0	14
7	45.5	8	57.0	12	56.5	11	39.5	10	29.5	13
8	51.5	3	78.5	4	61.0	8	36.0	11	42.5	7
9	46.0	7	72.0	7	66.0	4	48.0	3	44.0	6
10	41.0	10	68.0	9	61.0	8	45.5	4	41.0	9
11	59.0	1	83.5	1	65.5	5	51.0	1	47.0	4
12	44.0	9	59.5	10	55.0	12	***	***	40.0	10
13	46.5	6	57	12	67.5	3	41.0	8	48.0	3
14	51.5	3	74.5	5	57.5	10	42.5	5	47.0	4

* 120 pts. possible ** 80 pts. possible *** did not participate

metaevaluation are presented in conjunction with the three procedures used during the metaevaluation.

Metaevaluation Procedure - Evaluator Self-Report

The evaluator who designed and conducted the evaluation of the September 1981 session prepared a self-report that described the problems encountered during the fieldtest of the evaluation framework. The self-report was prepared in direct response to the research question #1.

1. What specific problems were encountered in the fieldtest of the evaluation framework?

The major problems with the fieldtest related to the collection of behavioral data. Originally the plan was to evaluate the performance of the participants in three different videotaped activities. However, the content and length of the activities, a clinical teaching simulation, a practice teaching assignment, and the presentation of a proposed research or evaluation project, proved to be too dissimilar to evaluate with a common rating scale. The clinical teaching simulation focused on skills related to supervising medical students and providing constructive feedback. The participants were allowed to choose their topic for the practice teaching assignment and their selections ranged from giving lectures to skill teaching to leading a committee meeting. The third activity was an opportunity to practice presentation skills.

As a result, an alternate design was formulated using two videotaped presentations. One of the presentations was the third activity identified above. The second was a videotape of the participants presenting the results of projects they had conducted. Since both activities were presentations, it was possible to develop a rating scale and items that could be used with both.

The original plan would have collected data at approximately the same times during the fellowship as the pretest, posttest, and delayed posttest. The new plan resulted in the collection of behavioral data in January and May. Thus, the types of judgments that could be made about the fellows using behavioral data had to be altered and the corresponding evaluation questions were revised.

Once the decision was made to rate the two presentations, the development of the rating scale was greatly facilitated. Rating required a considerable amount of time, but the rating task was facilitated by the decision to rate two similar presentations.

Other problems encountered during the fieldtest related to the development and administration of the cognitive pretest, posttest, and delayed posttest. Difficulties arose in the development of the test items. The time allotted for test development was insufficient to allow the test to be tested on a suitable group prior to its use with the fellows. Seven individuals submitted items of varying format and difficulty and even though the evaluator and program directors rewrote many of the items, the test was still not as well constructed as it might have been.

Several factors related to the administration of the test created additional problems. The pretest was administered during the initial orientation session and the 40 short answer, essay questions on the test took longer to complete than the evaluator and program directors had expected. The excessive amount of time required was attributed to the type of test items as well as to the content of the test which was new to the fellows. The fellows felt compelled to respond to as many items as possible and thus spent more time completing the test than expected. To

avoid the loss of further session time, the posttest and delayed posttest were administered as take-home tests with explicit instructions to treat them as closed-book tests to be completed alone.

In addition to the procedural problems associated with the cognitive tests, the validity of the test was a concern for all three administrations of the test. Although the participants' scores improved substantially from pretest to posttest and dropped slightly from posttest to delayed posttest, the percentage scores were lower than expected. Even on the subscales, a score of 76% represented the highest average score.

The test items were written by the faculty responsible for the presentation of that content. Although the evaluator and program directors edited the items, the content represented by the items still reflected the content of the presentation. A close examination of the test items and handouts for corresponding presentations verified this statement. However, while the content of the September session was oriented to practical applications of skills and techniques to teaching, many of the test items did not require applications or generalizations of content of the session to other situations. Rather the items required rote recitation of facts or lists or other comprehension-level activities. Thus the practical focus of the presentations was not measured by many of the test items. The application of the content of the session was not assessed by the cognitive test.

In addition to the examination of the test items and handout materials, an item analysis was conducted for each item on each administration of the test. A difficulty index was calculated for each item and each test. "The difficulty (or easiness) of a test item is

determined by the percent (P) of students who answered correctly" (Kryspin & Feldhusen, 1974, p. 136). Thus a difficulty index of .60 would indicate that 60% of the students answered that test item correctly. Kryspin and Feldhusen provided guidelines to use when assessing item difficulty levels. The guidelines are summarized in Table 17.

TABLE 17

DIFFICULTY LEVELS OF TEST ITEMS

DIFFICULTY INDEX	DIFFICULTY LEVEL
0 - .25	Hard
.26 - .74	Average
.75 - 1.00	Easy

Kryspin and Feldhusen also suggested the average item difficulty on a test should be between .50 and .70. Mehrens and Lehmann (1978) reported that the ideal average difficulty for a maximally discriminating completion and short answer test was .50.

The discrimination index indicates how well the students who performed well on the test performed on an item in relation to the performance of students who scored poorly on the test. If a test item truly discriminates, then students who perform well on the test perform better on the item than do students who perform poorly on the test. A negative discrimination index indicates that more students near the bottom of the ranking fared better on the item than did students ranked near the top of the group. Kryspin and Feldhusen proposed the guidelines in Table 18 for assessing discrimination indices.

TABLE 18

DISCRIMINATION LEVELS OF TEST ITEMS

DISCRIMINATION INDEX	LEVEL OF DISCRIMINATION
.40 - 1.00	Item discriminates well
.20 - .39	Item discriminates moderately well
.00 - .19	Item discriminates poorly
-1.00 - .00	Item discriminates negatively and needs revision or rejection

Difficulty and discrimination indices should be considered together when analyzing test items. When the difficulty and discrimination indices for the items on all three tests were calculated, seven poor items were identified. These items were "hard" items on at least two of the tests and five of the items were also negative discriminators. Three of the items were within the elements of group development subscale, the other four were distributed throughout the test.

The overall test results also indicated that the the tests poorly discriminated among the top and bottom groups of participants. The difficulty and discrimination indices for the tests are presented in Table 19.

TABLE 19

TEST DIFFICULTY AND DISCRIMINATION INDICES

INDEX	PRETEST	POSTTEST1	POSTTEST2
Difficulty	.38	.57	.50
Discrimination	.10	.15	.12

A positive finding relative to the test analysis was that 22 of the 40 items had difficulty indices of .50 or greater on at least two tests.

Since a difficulty index of .50 was considered average, this finding indicated that over 50% of the items were of average or easy difficulty. However, the discrimination indices were still low enough to doubt the overall quality of the test items and the test.

In conclusion, test analysis results indicate that at least seven of the test items on the cognitive test were of little or no value. These seven items in combination with the test's poor discrimination indices were sufficient reason to question the validity of the results.

The only other aspect of the fieldtest that presented any substantial difficulties were the telephone interviews with the fellows and supervisors. Several interviews were interrupted, shortened or re-scheduled due to patient care emergencies or medically related interruptions that occurred while the supervisor or participant was being interviewed. During participant interviews the evaluator frequently reminded the participants which topics were covered in September and which faculty presented which topics. Some participants had difficulty answering certain questions because they were not teaching as part of their current jobs. With supervisors, the problems related to their amount of contact with participants on issues or activities related to teaching. These problems occurred during interviews with individuals in both groups of interviewees, but were not characteristic of each individual interviewee.

Issues related to the internal and external validity of the fieldtest evaluation process were addressed during the fieldtest. Campbell and Stanley (1966) identified eight classes of extraneous variables which should be controlled if internal validity of experiments is to be maximized. Five of the eight variables were controlled during

the fieldtest by the evaluation design and procedures. These five were instrumentation, statistical regression, selection, mortality, and selection-maturation interaction.

The design and the practical constraints of the fieldtest made it difficult to control the remaining three variables. The evaluation of the session was designed to be as unobtrusive as possible in order to minimize disruptions of the fellows and session activities.

History was a problem variable because it was difficult to determine what additional events might have occurred between the posttest and delayed posttest to account for changes in test scores. Interview questions related to additional study and handout use addressed the issue of history, but did not completely control the effects of the variable. There was no suitable mechanism to control for the effect of maturation of the fellows over the nine months during which data were collected. Since the same test was administered three times, testing certainly had an effect on the internal validity of the fieldtest.

While these three classes of variables may have affected the internal validity of the fieldtest, the steps required to control the variables would have significantly reduced the relevance of the evaluation and most certainly would have made the measures more obtrusive. A compromise was made in the design of the fieldtest and as a result history, maturation, and testing were not controlled.

The external validity or representativeness of the fieldtest results was acknowledged to be extremely low from the beginning. Due to the design used for the fieldtest, there was little or no expectation that results could be generalized to other populations. Indeed, generalization of the results of the fieldtest was not a purpose of the study.

To summarize, the major problems with the fieldtest related to the collection of cognitive and behavioral data. Some changes were made in the orginal evaluation design to facilitate the design of instruments and the administration of the instruments. The development and administration of the tests required considerable amounts of time, more than expected. The validity of the test results was questioned. There were logistical constraints present during the telephone interviews and some data were difficult to gather due to the respondents' inability to remember or lack of teaching responsiblities. None of these problems caused major changes in the conduct of the evaluation and the fieldtest was conducted according to plan except for the previously described alterations. The internal validity of the fieldtest was compromised to maintain relevance and reduce disruptions to the session. External validity of the fieldtest results was not a goal of the fieldtest.

Metaevaluation Procedure - Interview of the Program Directors

The second metaevaluation procedure collected information to answer the remaining four research questions. For each research question, one or more evaluation standards from Standards for Evaluations of Educational Programs, Projects, and Materials were selected to be applied to the results and procedures of the fieldtest. Using the standards as guidelines, more specific questions related to the research questions were formulated. The specific questions were assembled as a questionnaire administered to the two program directors during a telephone interview conducted by the evaluator. The responses to these questions served as the basis for formulating responses to research questions #2, 3, 4, and 5.

A summary of the responses of the program directors to the specific questions related to each research question are presented in tabular form to facilitate data presentation. A transcription of the complete interview with the program directors is supplied in Appendix O. Additional questions were asked during the interview that were not related directly to any of the research questions. However, these questions related to how well the evaluation framework functioned in the fieldtest and the responses to these questions have implications for the study.

The research questions, evaluation standards, specific questions, and summarized responses of the program directors are presented in Tables 20, 21, 22, and 23. The program directors' responses are edited and combined for ease of representation. The responses to the questions not related directly to any of the research questions are provided in Table 24.

Metaevaluation Procedure - Evaluation Procedure Analysis

During the final stage of the metaevaluation, the procedures used during the fieldtest were evaluated by the evaluator and the program directors. Each of the procedures and the data collected with that procedure were rated using a three-point scale of high, medium, and low. The program directors also explained the rationale behind each of their ratings. The results of the ratings are displayed in Table 25 and Table 26.

The ratings of the six fieldtest evaluation procedures in relation to the five factors clearly delineated which procedures functioned well from those which functioned poorly. The End-of-Week Evaluations and final debriefing ranked first and second with the other four clustered together. The videotape rating ranked third and ranked behind only the

TABLE 20

RESPONSES TO RESEARCH QUESTION #2

RESEARCH QUESTION #2: Was the evaluation framework practical in its use of resources?

EVALUATION STANDARDS: The evaluation procedures should be practical, so that disruption is kept to a minimum, and that needed information can be obtained.

The evaluation should produce information of sufficient value to justify the resources expended.

SPECIFIC QUESTIONS	PROGRAM DIRECTORS' RESPONSES
1. Did the evaluation procedures produce information of sufficient value to justify the resources expended?	We felt justified in using the resources that we did for a one-shot deal. It would not be worth repeating all of those procedures every year, but they did provide some valuable data that we just didn't have.
2. Were the evaluation procedures administered so that program disruption was kept to a minimum?	They did not intrude much at all. The only possible exception is the pre-, post-, and delayed posttest. You did a lot of things like the telephone calls outside the program, but the procedures overall were not that invasive.
3. Did the use of multiple instruments appear to yield results that justified the extra time and effort involved in their development and administration?	Because it was a more comprehensive study, it was definitely justified the initial time. If everything is just confirming what you hear over and over again, you can start to select out that procedure that gives you the best information for the least effort. With the phone calls, the debriefing, and the End-of-Week Evaluations we were starting to get some redundancy in our information. If we were to do it again, we would probably chop out parts of that.

TABLE 21

RESPONSES TO RESEARCH QUESTION #3

RESEARCH QUESTION #3: Was the evaluation framework useful in providing information to the decision makers?

EVALUATION STANDARD: Information collected should be of such scope and selected in such ways as to address pertinent questions about the object of the evaluation and be responsive to the needs and interests of specified audiences.

SPECIFIC QUESTIONS	PROGRAM DIRECTORS' RESPONSES
1. Did the evaluation framework provide information that answered specific questions that you had about the program?	Yes, it did. Now we had data that told us that we were doing the right things or that we were off-base here. That was very helpful to us.
2. Was the information that you received complete and comprehensive? Was there anything left out that you would like to have known?	Perhaps the only information that was missing was some documentation of the fellows using the skills in their setting. They told us they do lectures and work with medical students, but if we could have actually observed them performing some of the skills in their own environment that would have been very helpful. There were times we felt inundated with information. For the individual information that was useful, but we might have wanted to see something a little more summarized.
3. How could the evaluation have been changed to provide more useful information?	Changing the cognitive level of the test items, making them more application and problem-solving rather than recall. Another possibility would be to get some baseline data on the performance skills, e.g., how well could they do presentations before they started, how well could they do clinical teaching before they started? Another thing that would be ideal, but would also put constraints on the program would be to standardize the presentations so that they were more equivalent to begin with. It

TABLE 21 (Cont'd.)

SPECIFIC QUESTIONS	PROGRAM DIRECTORS' RESPONSES
	would be easier for us to determine if they had really been applying what we have been talking about. At the same time the fellows would lose some of the freedom of picking what they want to do.

TABLE 22

RESPONSES TO RESEARCH QUESTION #4

RESEARCH QUESTION #4: Were the methods and instruments used during the fieldtest of the evaluation framework technically adequate?

EVALUATION STANDARDS: The sources of information should be described in detail so that the adequacy of the information can be assessed.

The information-gathering instruments and procedures should be chosen or developed and then implemented in ways that will assure that the interpretation arrived at is valid and sufficiently reliable for the intended use.

Quantitative and qualitative information in an evaluation should be appropriately and systematically analyzed to ensure supportable interpretations.

The conclusions reached in an evaluation should be explicitly justified, so that the audience can assess them.

SPECIFIC QUESTIONS	PROGRAM DIRECTORS' RESPONSES
1. Were the sources of information described in enough detail for you to assess the validity of the information they provided?	Yes.
2. Were the information-gathering instruments described in enough detail for you to assess the validity of the results they produced?	Yes, because we're familiar with it, but we don't know how someone who isn't intimately familiar with the program would respond.

TABLE 22 (Cont'd.)

SPECIFIC QUESTIONS	PROGRAM DIRECTORS'S RESPONSES
3. Were the information-gathering instruments described in enough detail for you to assess the reliability of the results they provided?	There's a lot of room just in the design for unreliability, but at the same time that's something you have to live with. You made a good effort at removing as much of the unreliability as you can.
4. Was there evidence that the data were collected and analyzed systematically?	Yes.
5. Did it appear that the quantitative data were appropriately and systematically analyzed?	Yes, you made a good effort at analyzing it.
6. Did it appear that the qualitative data were appropriately and systematically analyzed?	Yes, as much as you can subject any of that data to analysis. It tends to get a little voluminous at times. That's where we got inundated with information.
7. Were the conclusions presented in the evaluation report supported by the data?	

TABLE 23

RESPONSES TO RESEARCH QUESTION #5

RESEARCH QUESTION #5: Were the methods and instruments used during the fieldtest of the evaluation framework conducted in an ethical manner?

EVALUATION STANDARDS: Oral and written evaluation reports should be open, direct, and honest in their disclosure of pertinent findings, including the limitations of the evaluation.

Evaluations should be designed and conducted, so that the rights and welfare of the human subjects are respected and protected.

SPECIFIC QUESTIONS	PROGRAM DIRECTORS' RESPONSES
1. Was the evaluation report open, direct, and honest in its disclosure of pertinent findings, including the limitations of the evaluation?	Yes.
2. Was the evaluation designed and conducted so that the rights and welfare of the human subjects were respected and protected?	You were up front at all times with them and there was no hidden agenda.

TABLE 24

ADDITIONAL METAEVALUATION QUESTIONS AND RESPONSES

GENERAL QUESTION: How well did the evaluation framework function when applied to the Program?

SPECIFIC QUESTIONS	PROGRAM DIRECTORS' RESPONSES
1. What would you do differently if you were to do it again?	It is a job for someone other than the director or assistant director. We would use the End-of-Week Evaluations, the cognitive pre-, post-, delayed posttest, although we would want that to be a different type of cognitive measure. We would do the videotape presentations and ratings and the final debriefing. It's the interviews that we have some questions about in terms of the cost, the time, and we think we can get the data from other sources.
2. What was not done that should be done?	We could incorporate the two visits we make to a setting, the one for the pre- and the one during fellowship into a data collection effort. We could use that opportunity to somehow collect data.
3. Which evaluation source provided the information of most value to you?	The participants.
4. Which evaluation method provided the information of most value to you?	The ratings of the actual performance, the presentations. The basic question that we were interested in was "To what degree is what we're doing transferrable?" In other words, what can they now put into operation. That data helped us a great deal. In terms of overall acceptability, the End-of-Week Evaluations are what we look at. If we had to pick two instruments working together, one where we directly observe their performance and one where they give us information on how well we did on our presentations, the two together would be the best guess.

TABLE 24 (Cont'd.)

SPECIFIC QUESTIONS	PROGRAM DIRECTORS' RESPONSES
5. Which type of data would you rely on?	Performance data.
6. Which type of data would you least rely on?	The cognitive data and the supervisors' report.
7. What was the overall strength of the evaluation?	The strength is that it did make an attempt to collect data about both cognitive and affective outcomes and it used a variety of different sources of data, different data collection methods, and there was some redundancy and that was good in this case. You managed to hone in on what are the three best types of information. The weakness was that in an attempt to be comprehensive, we might have taken in too much information. Maybe we had too many measures under each of those different types of data.
8. What are you doing this time?	It may be a little early to tell. We haven't implemented any of the additional things that you've done yet, primarily because our questions were answered. We have not made any decisions if we're going to do any of the follow-up interviews yet, but it may affect how we gather information when we go out to do the site visits. We videotaped them, no thought yet of doing any analysis. It's not because we don't feel it's important, but this is the fifth year of the program and we have been doing the same thing for three years and your evaluation came in the fourth year and provided confirmatory data. We discovered that some of our initial assessment of their performance was substantiated with fairly rigorous ratings. We were not that far off.

TABLE 25

INDIVIDUAL RATINGS OF EVALUATION PROCEDURES

	FACTOR														
	DIRECT UTILITY			TIME EFFICIENCY			RESOURCE EFFICIENCY			CREDIBILITY			DATA MANAGEABILITY		
PROCEDURE	E	A	B	E	A	B	E	A	B	E	A	B	E	A	B
End-of-Week Evaluations	3	3	3	3	3	3	3	3	3	3	3	2	3	3	3
Cognitve tests	2	1	1	1	1	1	1	1	1	1	2	1	3	2	2
Videotape ratings	3	2	3	1	1	1	1	1	1	3	2	3	3	2	2
Fellow interviews	2	2	2	2	2	1	2	2	1	2	2	2	2	1	1
Supervisor interviews	2	1	1	2	2	1	2	2	1	2	1	1	2	1	1
Final debriefing	3	3	2	3	3	2	3	3	2	3	2	2	2	2	2

KEY: E - evaluator
A - program director A
B - program director B

SCALE: 3 - high
2 - medium
1 - low

TABLE 26

MEAN OVERALL RATINGS OF EVALUATION PROCEDURES

PROCEDURE	FACTOR: DIRECT UTILITY	TIME EFFICIENCY	RESOURCE EFFICIENCY	CREDIBILITY	DATA MANAGEABILITY	TOTAL
End-of-Week Evaluations	3.0	3.0	3.0	2.7	3.0	2.9
Cognitive tests	1.3	1.0	1.0	1.3	2.3	1.4
Videotape ratings	2.7	1.0	1.0	2.7	2.3	1.9
Fellow interviews	2.0	1.7	1.7	2.0	1.3	1.5
Supervisor interviews	1.3	1.7	1.7	1.3	1.3	1.5
Final debriefing	2.7	2.7	2.7	2.3	2.0	2.5

SCALE: 3 - high
2 - medium
1 - low

End-of-Week Evaluations in terms of the utility and credibility factors. Low rating on time and resource efficiency brought down the overall rating for the videotape rating.

The cognitive tests were rated low on all factors except data manageability. The interviews of the fellows and supervisors rated medium or below for all factors. The data collected with this procedure are summarized with the information gathered with the other two procedures in the following summary of the results of the metaevaluation.

SUMMARY OF RESULTS OF METAEVALUATION

The three metaevaluation procedures successfully identified the strengths and weaknesses of the fieldtest of the evaluation framework. These strengths and weaknesses are addressed in major findings presented in this summary. A number of the previously presented results also have implications beyond the study and are discussed in Chapter Five. Table 27 contains the five research questions and a brief answer to each question. The answers are based on the results of the metaevaluation.

The additional data gathered during the metaevaluation helped identify those procedures and data that were most useful to the program directors. The three preferred procedures were the End-of-Week Evaluations, final debriefing, and videotape ratings. The cognitive tests were rated lowest due to the question of the validity of the test data.

Among the three types of data collected, the program directors identified behavioral data as being more useful to them than cognitive and reaction data. However, the program directors indicated that a strength of the fieldtest of the evaluation framework was that reaction, cognitive, and behavioral data were collected using a variety of sources

TABLE 27

SUMMARY OF RESPONSES TO RESEARCH QUESTIONS

RESEARCH QUESTION	RESPONSE
1. What specific problems were encountered in the fieldtest of the evaluation framework?	The major problems related to the collection of behavioral data and to the development, administration, and validation of the cognitive tests.
2. Was the evaluation framework practical in its use of resources?	Yes, the program directors felt justified in committing the resources required to conduct the fieldtest.
3. Was the evaluation framework useful in providing information to the decision makers?	Yes, the data were comprehensive and confirmed the program directors' subjective assessments of the program's quality and impact.
4. Were the methods and instruments used during the fieldtest of the evaluation framework technically adequate?	Yes, with the exception of the cognitive tests. There were reasons to question the validity of the cognitive test results.
5. Were the methods and instruments used during the fieldtest of the evaluation framework conducted in an ethical manner?	Yes, the methods and instruments were conducted in an ethical manner. The evaluator was candid in his interactions with people during the fieldtest.

of information and collection methods. A weakness of the fieldtest was that too much of the information collected was redundant and at times overwhelming in its volume. This volume of information was attributed to the number of methods used and to the high proportion of open-ended questions included on many of the instruments.

In conclusion, the results of the metaevaluation demonstrated that the evaluation framework, as configured in the fieldtest, was successful in evaluating the impact of the September 1981 session of the Program. The program directors received valuable information used to make decisions about the program and the fieldtest produced results indicating

that cognitive and behavioral change had occurred in the group of fellows.

SUMMARY OF THE CHAPTER

The data gathered during the study were presented in this chapter. First the results of the fieldtest were provided. These results were paired with the appropriate evaluation questions, which were also grouped according to the type of data gathered, reaction, cognitive, or behavioral. Tables were used to summarize data when appropriate and summaries were provided for each data type. Finally, a summary of the fieldtest results was presented.

The results of the metaevaluation of the fieldtest of the evaluation framework were also provided in this chapter. A self-report prepared by the evaluator highlighted problems encountered during the fieldtest and served as a response to the study's first research question. Responses to the other four questions were furnished based on information collected during an interview with the program directors. The responses to additional questions asked during the interview were also supplied. The results of a third metaevaluation procedure, the rating of fieldtest evaluation procedures and results, were also presented. Finally, a summary of the results of the metaevaluation was provided.

In the concluding chapter of the dissertation, a summary of the first four chapters is provided. The results of the fieldtest and metaevaluation are discussed and conclusions are drawn. In closing, recommendations are made for further research and implications of the study for educational practice are considered.

CHAPTER FIVE

SUMMARY AND CONCLUSIONS

INTRODUCTION

In this chapter the study is summarized. The problem, literature, procedures, and results of the study are reviewed. The study results are discussed and conclusions are drawn. In conclusion, recommendations for further research are suggested and implications of the study for educational practice are considered.

THE PROBLEM

Short-term training is a popular format used in training and education throughout the United States. One specific purpose for which short-term training programs are frequently used is to improve the teaching skills of faculty in post-secondary education. Faculty development programs exist in a number of colleges and universities, particularly medical schools.

The problem that led to this study was that few faculty development programs evaluate their effectiveness. The majority of existing faculty development programs relied on participant self-reports and reaction data to measure the effect of faculty development activities. Evidence of

cognitive or behavioral change in the participants was rarely reported. In addition, no evaluation approach designed specifically for short-term training programs was identified. In response to the problem, the study reported in this dissertation was designed and conducted to determine whether an evaluation framework for short-term training programs could be developed and successfully implemented.

THE LITERATURE

The literature of the evaluation of faculty development activities in post-secondary education and medical education was characterized by a dependence on self-reports and satisfaction data. Numerous authors, including Centra (1976), Stephens (1981), and Levinson and Menges (1979), stressed the importance of gathering objective data on cognitive and behavioral change to assess program impact on participants.

The review of the literature on evaluation models confirmed the initial assumption that no single evaluation model existed that was suited to the task of evaluating short-term training. However, numerous concepts and components were identified within existing models that were applicable to an evaluation framework for short-term training programs.

The section of the literature review that focused on evaluation methodology contained an inspection of several critical evaluation procedures. The evaluation procedures examined included the relative value of using quantitative and qualitative methods and the rationale for using multiple methods, multiple sources of information, and multiple levels of evaluation. Many of the concepts discussed in this section were incorporated into the design of the evaluation framework.

The final section of the literature review focused on the process of metaevaluation, a relatively new concept. Very few authors have contributed to the field of metaevaluation and the literature was correspondingly sparse. However, suggested approaches were presented and one approach became the framework for the metaevaluation conducted in this study.

PROCEDURES AND METHODS

In Chapter Three, the procedures used in the design, development, and fieldtest of the evaluation framework were described. The major components of the framework were detailed and displayed in a matrix. The program evaluated during the fieldtest was described and a matrix depicting the evaluation plan for the fieldtest was introduced. Evaluation instruments employed during the fieldtest were described and data analysis procedures were outlined. The chapter concluded with a description of the procedures used in the metaevaluation of the fieldtest.

RESULTS

Fieldtest results and outcomes of the metaevaluation were presented in Chapter Four. Results of the evaluation of the September 1981 session of the Family Medicine Faculty Development Program (Program) were reported with the 18 evaluation questions that guided the evaluation. Metaevaluation data were paired with the study's research questions and the evaluation standards. The major findings drawn from these results were presented and discussed.

DISCUSSION

Issues related to the results of the study are discussed in this section of the chapter. The major issues include reaction data, cognitive data, behavioral data, evaluation procedures, and the evaluation framework. The five issues are considered in relation to the results of the fieldtest and metaevaluation and to issues discussed in previous chapters of the dissertation.

The results of the fieldtest demonstrated the redundancy of the reaction data collected with various evaluation procedures from different information sources. The redundant data were reassuring information since the data collected with the End-of-Week Evaluations in September were supported by the data collected via the interviews in March and the final debriefing in May. The reaction data gathered from the program directors were consistent with the comments of the fellows in most instances and focused on the same weaknesses and strengths of the session.

The redundancy of information was noteworthy in that the reliability and validity of the data gathered during the End-of-Week Evaluations were strengthened by the data collected by the evaluator during the interviews. The reliability and validity of the data collected in September and March were further supported by the data collected during the final debriefing in May.

The reaction data were also notable in terms of their relationship with other outcomes of the study. There was a relationship between participant reactions to topics and presentations during the September session and their subsequent behavior and performance.

The terms "relevant" and "applicable" were frequently used by the fellows in their comments. The presentations the fellows enjoyed and perceived as most relevant or applicable to their present or future activities were also the subscales on which they had the highest scores on the cognitive tests. The reported use of handouts from these presentations and the reported use of skills and techniques related to these topics were also highest among all the topics and presentations. Participant comments related to the session on presentation skills were also supported by the videotape ratings, especially comments by the fellows pertaining to their improved organization while lecturing and teaching.

Among the cognitive results, the issue of the poor quality of the test was the primary concern. Certain logistical constraints, such as lack of time to pilot the test prior to its use with the fellows, were considered previously. Suggestions for improving the test are described below.

A major drawback of the cognitive test was that it did not test the participants' ability to apply or generalize the content of the September session to other situations. Over half the test items focused on recall of facts, lists, and definitions. The test could be improved substantially if more of the test items were rewritten to test application rather than recall of information.

Bloom, Hastings, and Madaus (1971) discussed evaluation techniques for assessing application of instruction.

> Teachers and curriculum makers have long recognized that a student doesn't really "understand" an idea or principle unless he can use it in new situations. Thus, application is frequently regarded as an indication that a subject has been adequately mastered. (p. 159)

Bloom et al. acknowledged the difficulty of developing items that measure the learner's ability to apply principles and generalizations to new problems and situations.

> The posing of new problems and situations is a difficult art in evaluation. It requires the evaluator to find or make new problems and situations within the grasp of his students. It is especially useful if the problems are real ones rather than contrived ones, with artificial or fictitious elements. Students find real problems more satisfying to attack than patently contrived problems, which can seem rather like puzzles and tricks to be solved. (p. 162)

Rules for generating application-level test items were suggested by Bloom et al. In short, they stressed the importance that the problem situation be new, unfamiliar, or somehow different from the situations used during instruction. The test item difficulty is determined in part by how different it is from the problems presented during instruction. The use of appropriate principles of generalization should be required to answer the test questions. The final rule suggested by Bloom et al. was that one or more of the following behaviors should be sampled by each test item.

> The student can determine which principles or generalizations are appropriate or relevant in dealing with new problem situations.
>
> The student can restate a problem so as to determine which principles or generalizations are necessary for its solution.
>
> The student can specify the limits within which a particular principle or generalization is true or relevant.
>
> The student can recognize the exceptions to a particular generalization and the reasons for them.
>
> The student can explain new phenomena in terms of known principles or generalizations.
>
> The student can predict what will happen in a new situation by the use of appropriate principles or generalizations.

> The student can determine or justify a particular course of action or decision in a new situation by the use of appropriate principles or generalizations.
>
> The student can state the reasoning he employs to support the use of one or more principles or generalizations in a given problem situation. (p. 165)

The rules proposed by Bloom et al. for developing application-level test items could be used to improve the cognitive test used during the fieldtest. Recall items could be rewritten or new items could be generated. A positive outcome of developing application-level items is that poor results may be attributed to the learner's failure to learn the material rather than the uncertainty induced by the questionable validity of the test items on the fieldtest cognitive tests.

The other issue discussed in relation to the cognitive results of the fieldtest is the participants' reports of additional study and subsequent handout use. The most additional study of a topic reported by the fellows was less than half of the group, 6 of 14. The least additional study reported was 1 of 14. Several explanations are considered during the following discussion of the lack of additional study.

One explanation for the lack of additional study is that the fellows did not have the time or motivation to do extra reading or attend additional workshops on the topics of the September session. Conversely, the fellows may not have wanted to pursue additional study because they had learned what they needed to know about the topics and further study was not required. This issue was not resolved by the study, but additional study was not reported by a sufficient number of fellows to suspect that additional study had an effect on the retention of knowledge as measured by the delayed posttest.

Handout use was more pronounced than additional study and may have affected delayed posttest scores. However, the topics for which handouts were used the most were also the topics which introduced the skills and techniques most frequently used by the fellows. The possibility that the delayed posttest scores were affected by referral to handouts was counterbalanced by the fact that the purpose of the September session was to provide the fellows with practical skills and techniques they could use while teaching. If no additional study was required due to participant satisfaction with the content of the session and if handouts were used to prepare for specific teaching activities, then the session was successful in a manner which the delayed posttest could not measure.

Several issues related to behavioral data collected during the study deserve discussion. The first issue is the videotape rating scale used to measure the fellows' ability to apply presentation skills in two videotaped presentations.

The videotape rating scale was developed specifically to assess the presentation skills of the participants in the September 1981 session of the Program. Handout materials from the sessions on principles of learning and motivation and presentation skills provided the basis for the content of the 16 items on the videotape rating scale form. Unlike the cognitive test results, the videotape ratings were of high validity and provided the program directors objective information related to the fellows' strengths and weaknesses as presenters. As a result, the videotape rating scale was much more likely to be implemented by the program directors in the future than were the cognitive tests or interviews conducted during the fieldtest.

The issue of self-report data is considered next. It was suggested in the faculty development literature that self-reports should not be relied upon as the sole evidence of the effectiveness of faculty development programs. Centra (1976), Stephens (1981), and Levinson and Menges (1979) were critical of the use of self-report data in the absence of cognitive or behavioral measures of faculty improvement as teachers. Yet in most instances reported in Chapter Two, reaction data and/or self-reports were used to measure impact of instructional improvement activities.

Distrust of self-report data as evidence of behavioral change is not limited to the faculty development literature. Howard, Schmeck, and Bray (1979) reported:

> It is axiomatic that, given a choice between a self-report and a behavioral measure of the same phenomenon, researchers will choose the behavioral measure. Likewise, when behavioral and self-report indices of the same construct show substantial discrepancies, it is seen as a signal to suspect the self-report measure rather than the behavioral measure. (p. 129)

Howard, Maxwell, Wiener, Boynton, and Rooney (1980) added:

> The status of self-report techniques in modern research is clearly that of a second-class citizen. Critiques of self-report approaches, representing detours on the road to a truly rigorous scientific discipline, are ubiquitous.... Researchers are advised to employ self-reports only if no behavioral index of a construct exists, such as with dogmatism, or if behavioral measures are too difficult or too costly to obtain. (p. 293)

However, recent research has been conducted to determine whether the validity of self-report data can be increased. A Retrospective Pretest-Posttest Design, similar to the approach used during the interviews with the fellows to determine expertise in the session topics and performance in the three simulations or presentations, was proposed by Howard et al. (1979).

> This design would simply improve a modification of Campbell and Stanley's Design 4 to include a retrospective pretest at the time of posttesting. This is accomplished by asking subjects to respond to each item on the self-report measure twice. First, they are to report how they perceive themselves to be at present (Post). Immediately after answering each item in this manner, they answer the same item again, this time in reference to how they now perceive themselves to have been just before the workshop was conducted (Retrospective Pre). Subjects are instructed to make the Retrospective Pre responses in relation to the corresponding Post response in order to insure that both responses are made from the same perspective. Each set of ratings is scored separately to yield a Post score and Retrospective Pre score. (pp. 130-131)

The results of a Retrospective Pretest-Posttest Design are still not conclusive, but the selection of the design is an option available to evaluators. Howard et al. (1980) stated, "The present set of studies demonstrates that some of the evidence traditionally cited to demonstrate the lack of accuracy of self-reports must be reconsidered" (p. 309). Thus, the evaluator forced to rely on self-reports to gather a portion of the data to evaluate a short-term training program should consider a Retrospective Pretest-Posttest Design to increase the accuracy of self-report data.

The issue of accuracy of self-report data is also of importance when considering certain discrepancies between the subjective and objective data gathered during the fieldtest. In particular the data gathered for three fellows are discussed, since the program directors indicated their disappointment with their performance. Overall test results, presentation skills subscale results, videotape ratings, program director comments, fellow self-reports, and supervisor comments are described. The fellows' rankings are presented in Table 28.

With the exception of Fellow 6, who ranked near the top of the group on the posttest and delayed posttest, the objective data indicated the

three fellows ranked near the bottom of the group on all five measures, including the presentation skills subscale. This finding was consistent with the program directors' assessment of the three individuals.

> We were disappointed in their skill and motivation level.... This group couldn't apply information, just rote recitation. We were disappointed in their ability to verbalize what we were trying to teach.... In terms of taking advantage of what the program had to offer, they played around with their projects, nothing will change in their lives as a result of being in our fellowship.

The comments made by the fellows and the supervisors depicted a different view of the skills and abilities of the three fellows. The fellows all reported an increase in expertise in presentation skills from September to March in addition to predicting improved performance if the proposed research presentation were repeated. Use of the handouts related to presentation skills was reported by two of the fellows.

Supervisors of two of the fellows noted increased organization in the fellows' teaching. The third supervisor reported having had little personal contact with the fellow since September and was unable to judge changes in teaching skills.

An obvious discrepancy exists concerning the performance of three fellows. The comments of the participants and their supervisors do not coincide with objective test data, rating of two presentations, and comments of the program directors. One explanation is that the three individuals did improve in the area of presentation skills as applied in their home institutions. This would explain the discrepancy between data collected based on program activities and data collected based on activities on the job. Another explanation might be that the fellows' improvement during the program activities was so slight compared to other fellows that it was not detected by the program directors or objective

TABLE 28

RANKINGS FOR THREE SELECTED PARTICIPANTS

FELLOW	PRETEST OVER-ALL	PRETEST PRES. SKILLS	POSTTEST1 OVER-ALL	POSTTEST1 PRES. SKILLS	POSTTEST2 OVER-ALL	POSTTEST2 PRES. SKILLS	VIDEO1	VIDEO2
1	14	14	11	13	14	12	13	10
6	12	3	3	1	6	5	12	14
7	8	7	12	10	11	5	10	13

measures. Yet that improvement was noted by the fellows and the supervisors. A third explanation is that the fellows and supervisors told the evaluator positive change occurred because that was the expected outcome of the September session and results otherwise would reflect poorly on the fellow, supervisor, and/or home institution. The study does not provide conclusive answers to the discrepancy between objective and subjective data on the three fellows, but three possible explanations have been offered.

The final issue related to behavioral data concerns its relationship to cognitive data. This issue also has implications for the three fellows discussed previously. As Davis (1979) stated, "The fact that one has learned something does not in any way guarantee that the knowledge will manifest itself in a change in performance" (p. 125). One of the assumptions of short-term training is that if trainees learn new techniques, knowledge, and skills, the behavior or performance of the trainees on the job will change. Therefore it is assumed that cognitive learning must precede behavioral change or transfer of training. Ellis (1965) stated, "If we expect students to show much transfer in course work that involves general principles, we must be reasonably sure that the principles are thoroughly understood" (p. 72). Singer (1968) cautioned against considering learning and performance as being the same.

> Learning and performance are not synonomous terms, for performance is a function of an individual's past experience (learning) as well as other variables, mainly motivation. There are many occasions that, for various reasons, performance levels do not reflect the true amount of learning that has occurred. Although performance scores are the best indicators of learning as of the present time, an evaluation of learning should not be based solely on task performance. (p. 326)

The importance of the role of motivation in enhancing learning and subsequently performance was addressed by Ellis. "To the extent that

motivational variables influence learning, they are also likely to influence transfer. If a student is poorly motivated he will tend to learn less, thus reducing the chance of transfer to new learning situations" (p. 65). Singer also discussed the importance of motivation.

> Motivation distinguished performance from learning. If a person is unmotivated to perform the desired responses, he does not learn. Motivation is also a factor in explaining variable performances. (p. 332)

The relationship among learning, performance, and motivation is of great importance to this study. It has been demonstrated that noteworthy relationships existed among the fellows' overall scores on cognitive tests and videotape ratings and the fellows' subscale scores for topics of high and low relevance, use, and enjoyment. While the assumption is that learning new techniques and skills enhances future performance on the job, the motivation factor must be considered.

> In summary, the performance of the faculty member is determined by motivation and learning acting together. It is interesting to note that many faculty development programs tend to emphasize one or the other of these two variables, but not both. (Davis, 1979, p. 131)

Davis' comments have particular meaning for the results of the fieldtest since the three fellows discussed previously were all required to participate in the Program by their supervisors. For whatever reason, unlike 10 of the remaining 11 fellows, these fellows were not participants in the Program by their own choice. Perhaps their involuntary participation was reflected by a lack of motivation and hence by poorer performance in relation to the other fellows. Regardless of the issue of voluntary or involuntary participation, motivation of participants should be considered when assessing the outcomes of short-term training programs, especially as motivation affects initial learning of program

content and subsequent transfer of training and resultant behavioral changes.

A fourth major area to be discussed involves the evaluation procedures used during the fieldtest. Issues related to evaluation procedures include preferred procedures identified by the metaevaluation, the issue of qualitative versus quantitative data, and the volume of data generated by the various procedures used during the fieldtest.

The preferred procedures identified during the metaevaluation were the End-of-Week Evaluations, the final debriefing session, and the videotape ratings. The future use of a cognitive test was not precluded by the program directors, but the test used during the fieldtest would require extensive revision to be of value to the program directors. As mentioned previously, the videotape rating form was the only procedure designed specifically for the fieldtest that was identified as having immediate value to the program directors.

Quantitative and qualitative measures were used during the fieldtest. The End-of-Week Evaluations, fellow interviews, and supervisor interviews collected both quantitative and qualitative data. The cognitive test and videotape rating produced quantitative data only, while the final debriefing collected only qualitative data. If the preferred procedures identified in Chapter Four were used to re-evaluate the September session, both qualitative and quantitative data would be gathered. The primary concern, based on the outcome of the fieldtest and metaevaluation, would be to limit the number of open-ended questions which resulted in the overwhelming amount of evaluation data during the fieldtest. Since most of the open-ended questions were asked during the interviews, controlling the volume of data should not be a major problem.

The cognitive test data were manageable and concise and if the test is revised, the results would not be unwieldy or significantly add to the volume of data collected.

Patton (1980) provided some advice to evaluators which has implications for the results of this study.

> The evaluator using different methods to investigate the same program should not expect that the findings generated by those different methods will automatically come together to produce some nicely integrated whole. Indeed, the evidence is that one ought to expect initial conflicts in findings from qualitative and quantitative data, and expect those findings to be received with varying degrees of credibility. (p. 330)

Therefore, the discrepancies described in Chapter Four and in previous sections of this chapter are to be expected and should be shared with decision makers and other individuals interested in the results of the evaluation.

The final area of discussion focuses on the evaluation framework proposed, developed, and fieldtested in this study. Strengths and weaknesses of the framework are discussed. Included in the discussion is° consideration whether the strengths and weaknesses are attributed to the framework and its five components or to the procedures used within the framework during the fieldtest. Possible applications of the evaluation framework are also considered.

The strengths and weaknesses of the evaluation framework for short-term training programs are summarized in Table 29. Strengths or weaknesses attributed to the procedures used during the fieldtest are identified ccordingly.

TABLE 29

STRENGTHS AND WEAKNESSES OF THE EVALUATION FRAMEWORK

STRENGTH	WEAKNESS
Holistic, systems approach	Time required from implementation of evaluation framework to final results
Focus on outcomes and impact	No immediate formative evaluation feedback possible
Decision-oriented approach	Uncertainty whether all three types of data must be collected at all times (procedure)
Prescriptive, but options allowed when selecting information sources, data-gathering methods, and evaluation questions	Excessive amount of information collected (procedure)
May be used for formative or summative evaluation purposes	No explicit criteria delineated; left to discretion of program directors (procedure)
Allows use of qualitative and quantitative methods	Questionable validity of cognitive test (procedure)
Collects three types of data to provide comprehensive information to decision makers	
Collects redundant information (procedure)	

The strengths listed in Table 29 reflect the rationale presented in Chapter Three and require no further discussion at this point. However, the weaknesses that were identified during the study, both within the framework and the evaluation procedures, deserve elaboration.

The focus of the evaluation framework is on determining program impact on participants and their subsequent use of skills and knowledge learned during the short-term training program. Therefore, a major portion of the data collection must be conducted following the completion of the program. In the case of the fieldtest, data were collected as late as eight months following the session. No timelines were prescribed in the framework, but if retention of content and subsequent use of

program content is to be assessed, follow-up evaluation procedures have to be delayed to allow participants opportunities to integrate new skills and techniques into their daily routine. Therefore, final results of the evaluation are not available for use by the decision makers until after the completion of the program. For this same reason, if information is to be used for formative evaluation purposes, its use for making programmatic changes is delayed until subsequent offerings of the program.

The results of the study did not provide conclusive evidence concerning the necessity of collecting all three types of data proposed in Chapter Three. While the reaction and behavioral data were easily used by the program directors, it was difficult for them to use the cognitive test data for decision making given its low validity and poor ability to discriminate. Thus one cannot determine from the study if cognitive data in general would have utility for decision making. Therefore, the requirement to collect all three types of data is a weakness of the procedures used during the fieldtest.

The remaining weaknesses are also related to the procedures used during the fieldtest and have been discussed previously, particularly the cognitive test and excessive volume of data generated during the fieldtest. The lack of explicit criteria for the cognitive test and videotape rating results constitutes a weakness in the framework when results are examined by individuals external to the program. Without explicit criteria and cut-off scores, percentage scores have little or no meaning to other individuals examining the results of the fieldtest. However, this weakness is in the procedures used during the fieldtest, not in the evaluation framework.

The final issue to be discussed is the range of possible uses of the evaluation framework. The framework was designed specifically for one program type, short-term training. The framework was fieldtested on a short-term training program for family physicians interested in careers in academic medicine. To fully assess the value of the evaluation framework, it should be used with short-term training programs of varying length, content, and audience.

The possibility also exists that the evaluation framework might be as effective or more effective with programs that do not satisfy the characteristics of a short-term training program outlined in Chapter One. The first two weaknesses of the framework presented in Table 29 would appear to mitigate the framework's successful application to longer programs, such as college courses 10 to 15 weeks in length. However, no conclusive statements regarding use of the evaluation framework with programs that are not considered short-term training can be made based solely on the results of the fieldtest and metaevaluation.

In summary, a number of issues pertaining to the results of the fieldtest and metaevaluation were discussed. These issues were grouped into five major categories, reaction data, cognitive data, behavioral data, evaluation procedures, and the evaluation framework. Conclusions based on the issues discussed in this section are presented in the following section.

CONCLUSIONS

The following conclusions are drawn from the results of the study:

1. The evaluation conducted during the fieldtest of the evaluation framework collected information that documented the impact of the September 1981 session of the Program on 14 participants.

2. The metaevaluation identified three evaluation procedures which were considered most efficent and effective during the fieldtest.

3. Discrepancies in evaluation results should be expected when qualitative and quantitative data are gathered from a variety of information sources using different evaluation procedures.

4. The evaluation framework is not useful for the purpose of providing immediate formative evaluation information to decision makers.

RECOMMENDATIONS FOR FURTHER RESEARCH

The study described in this dissertation was not designed to allow a large number of generalizations to be made to other programs, populations, or settings. However, a number of the issues and results discussed in previous sections of this dissertation deserve further exploration. Recommendations for further research are based on issues that originated from the results of the fieldtest and metaevaluation.

As discussed in other sections of this chapter, the evaluation framework for short-term training programs should be applied to other programs, content areas, and populations. The purpose of applying the evaluation framework to other short-term training programs is to determine the optimal configuration and ultimate value of the evaluation framework. The weaknesses identified in Table 29 should be analyzed to determine if improvements can be made in the procedures or the framework. The identified strengths of the framework should be assessed to ascertain if they are considered strengths when the framework is applied to other programs.

The evaluation framework should be tested in direct comparison to other evaluation approaches. For example, a short-term training program should be evaluated simultaneously using the evaluation framework and one

of the other approaches discussed in Chapter Two. The outcomes of the two evaluations could be assessed and compared using procedures similar to the metaevaluation procedures described in the study.

Further research should be conducted to determine the optimal number and type of procedures necessary to evaluate short-term training programs. Issues surrounding the appropriate ratio of qualitative and quantitative measures to be used should be addressed. Is the Retrospective Pretest- Posttest Design proposed by Howard et al. (1979) effective with short-term training programs? What subjective measures can be trusted as accurate indicators of program outcomes? Is it necessary to collect all three types of data proposed in the framework? Each of these issues have possible ramifications for a more complete understanding of the evaluation of short-term training.

Additional research should consider assessment of participant motivation as a component of evaluating impact of short-term training programs. The issue of voluntary versus involuntary participation is an example of motivation factors that could be explored. Organizational factors could also be examined to detemine why utilization of short-term training program content is or is not occurring on the job.

Finally, the procedures developed for the fieldtest should be tested further. The videotape rating scale should be used and analyzed to assess its value for evaluating presentation skills with other programs and populations. The cognitive test should be revised to ascertain if it is possible to assess participant ability to apply program content using paper and pencil tests.

IMPLICATIONS FOR EDUCATIONAL PRACTICE

The ultimate value of the study described in this dissertation is based on the products, processes, and procedures which can be used by evaluators, educators, trainers, administrators, and other individuals responsible for short-term training. The procedures and instruments described in this dissertation could be adopted and modified for use in a number of situations. The conceptual approach to short-term training presented throughout the dissertation could be useful to practitioners. Use of the Retrospective Pretest-Posttest Design, consideration of participation motivation when assessing short-term training, and the use of behavioral rating scales and checklists to evaluate skills in practical situations are all possible practical outcomes of this study.

The primary purpose of the study was to develop an evaluation approach to assess the impact of short-term training programs. The evaluation framework was designed, developed, and fieldtested to serve that purpose. One particular short-term training program, the September 1981 session of the Family Medicine Faculty Development Program, was evaluated using an evaluation design based on the evaluation framework. The results of the fieldtest and metaevaluation of the fieldtest were presented in Chapter Four with the expectation that the results would be of interest and value to potential users of the framework.

Several conditions related to the Program require discussion because these conditions limit the general applicability of the results of the fieldtest. First, the group of 14 fellows were paid participants using release time, not vacation time, to participate in the fellowship. It is more often the case that the participants in short-term training programs or the participants' organizations pay to attend such programs.

Frequently, participants must use weekends, evening hours, or vacation time to attend these programs.

Second, the two-week segment evaluated during the fieldtest was part of a longer, continuing relationship between the fellows and the program directors. This continuing relationship facilitated certain data collection activities that might be more difficult to conduct under different circumstances. On several occasions, stipend checks were withheld until materials were submitted by fellows.

The evaluation framework was designed for use wherever short-term training may be provided. However, potential users of the framework should be aware of the effect the conditions described above may have had on the fieldtest results. Potential users should also consider how the absence or presence of similar conditions might affect further applications of the evaluation framework.

The literature review in Chapter Two indicated that the majority of the short-term training programs used for faculty development activities were evaluated solely on the basis of subjective participant reactions and self-reports. As discussed previously, several authors stressed the importance of gathering more rigorous data to objectively assess cognitive and behavioral change in participants in short-term training programs. This study demonstrated that behavioral data can and should be collected to assist decision making.

The framework used to collect the data in the fieldtest was designed eclectically, an approach recommended by Baron and Baron (1980), Patton (1980), and Steele (1973) in Chapter Two. The implication of their suggestions to follow an eclectic approach to evaluation was that the existing models were not functioning satisfactorily. Thus, the

evaluation framework for short-term training programs was designed using an eclectic approach and was fieldtested. The results were positive and the fieldtest was a success. The question of whether the evaluation framework is superior to other approaches remains unanswered and will remain unanswered until the evaluation framework is used by other individuals.

To fully assess the value of the evaluation framework, additional short-term training programs in other settings with different content, length, and populations must be evaluated with the evaluation framework for short-term training programs. Only then can more definite conclusions be drawn about the evaluation framework. As a conclusion to the dissertation and, hopefully, an introduction to further educational research and practice, the evaluation framework was modified to reflect the results of the metaevaluation and issues discussed in previous sections of this chapter. A revised matrix of the evaluation framework is outlined in Table 30. The revised matrix incorporates the major findings of the study into its design and offers an alternative to the matrix of the evaluation framework presented in Chapter Three.

TABLE 30

REVISED MATRIX OF THE EVALUATION FRAMEWORK

TYPE OF DATA	ASSESS-MENT OF	SOURCE OF DATA	METHOD OF GATHERING DATA	EVALUATION QUESTIONS
REACTION (SATISFACTION	STP	P	EOW,FD	How satisfied with the program were the participants?
COGNITIVE (LEARNING)	P	P	TESTS	How much of the content of the program did the participants learn and retain?
	P	P	S-R	How did the participants perceive their own learning and retention of the content of the program?
BEHAVIORAL (PERFORMANCE)	P	P	DO,VT	How well did the participants apply the content of the program in simulated or actual performance settings?
	P	P	S-R	How did the participants perceive their ability to apply the content of the program in simulated or actual performance settings?

KEY:
- STP - Short-term training program
- P - Participants
- EOW - End-of-Week Evaluations
- FD - Final debriefing
- S-R - Self-reports (using Retrospective Pretest-Posttest Design)
- DO - Direct observation
- VT - Videotape of participant performance in simulated or actual setting

SUMMARY

This dissertation presented the procedures and logic behind the design, development, and fieldtest of an evaluation framework for short-term training programs. The study described in this dissertation was conducted because there was a perceived void in the area of evaluation of impact of short-term training programs. The resultant evaluation framework was an attempt to partially fill that void. Initial findings from the fieldtest and subsequent metaevaluation indicate that the framework has the potential to be a valuable tool for those individuals responsible for planning, implementing, and evaluating short-term training programs. Additional applications of the framework to other-short term training programs will determine the ultimate value of this evaluation approach.

REFERENCES

LIST OF REFERENCES

Adams, W.R., Ham, T.H., Mawardi, B.H., Scali, H.A., & Weisman, R., Jr. Research in Self-Education for Clinical Teachers. Journal of Medical Education, December 1974, 49, 1166-1173.

Alkin, M.C. Evaluation Theory Development. UCLA CSE Evaluation Comment, 1969, (2), pp. 2-7.

Alkin, M.C. Towards An Evaluation Model: A Systems Approach. In P. Taylor & D. Cowley (Eds.), Readings in Curriculum Evaluation. Dubuque, IA: W.C. Brown Company, 1972.

Alkin, M.C., & Fitz-Gibbon, C.T. Methods and Theories of Evaluating Programs. Journal of Research and Development in Education, Spring 1975, 8 (3), 2-15.

Baron, J.B., & Baron, R.M. In Search of Standards. In R. Perloff & E. Perloff (Eds.), Values, Ethics, and Standards in Evaluation (New Directions for Program Evaluation, Number 7). San Francisco: Jossey-Bass, 1980.

Bell, N.T., Hunt, A., Parkhurst, P.E., & Tinning, F.C. Faculty Development in Michigan State University College of Osteopathic Medicine 1976-1978. Occasional Paper No. 6, East Lansing, MI: Michigan State University College of Osteopathic Medicine, 1979.

Bland, C.J. Faculty Development Through Workshops. Springfield, IL: Charles C. Thomas, 1980.

Bland, C.J., & Froberg, D.G. A Systematic Approach to Faculty Development for Family Practice Faculty. The Journal of Family Practice, 1982, 14 (3), 537-543.

Bland, C.J., Reineke, R.A., Welch, W.W., & Shahady, E.J. Effectiveness of Faculty Development Workshops in Family Medicine. The Journal of Family Practice, 1979, 9 (3), 453-458.

Bloom, B.S., Hastings, J.T., & Madaus, G.F. Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill, 1971.

Borg, W.R., & Gall, M.D. Evaluation Research: An Introduction (3rd ed.). New York: Longman, 1979.

Borich, G.D., & Jemelka, R.P. Definitions of Program Evaluation and Their Relation to Instructional Design. Educational Technology, August 1981, pp. 31-38.

Brethower, K.S., & Rummler, G.A. Evaluating Training. Improving Human Performance Quarterly, 1977, 5 (3-4), 103-120.

Britan, G.M. Experimental and Contextual Models of Program Evaluation. Evaluation and Program Planning, 1978, 1, 229-234.

Bryk, A.S. Evaluating Program Impact: A Time to Cast Away Stones, A Time to Gather Stones Together. In S.B. Anderson & C.D. Coles (Eds.), New Directions in Program Evaluation, Number 1. San Francisco: Jossey-Bass, 1978.

Caldwell, P.A. Preservice Training for Instructors of Adults. In S.M. Grabowski & Associates (Eds.), Preparing Educators of Adults. San Francisco: Jossey-Bass, 1981.

Campbell, D.T., & Stanley, J.C. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally, 1966.

Canfield, P.R. Family Medicine: An Historical Perspective. Journal of Medical Education, November 1976, 51, 904-911.

Centra, J.A. Faculty Development Practices in U.S. Colleges and Universities. Princeton, NJ: Educational Testing Service, 1976.

Cook, T.D., & Gruder, C.L. Metaevaluation Research. Evaluation Quarterly, February 1978, 2 (1), 5-51.

Corbett, T.C. Evaluation of Post-Implementation Strategies and Systems: Introduction. In P.P. LeBreton, V.E. Bryant, D.L. Zweizig, D.G. Middaugh, A.G. Bryant, & T.C. Corbett (Eds.), The Evaluation of Continuing Education for Professionals: A Systems View. Seattle: University of Washington, 1979.

Cronbach, L.J., Ambron, S.R., Dornbusch, S.M., Hess, R.D., Hornik, R.C., Phillips, D.C., Walker, D.F., & Weiner, S.S. Toward Reform of Program Evaluation. San Francisco: Jossey-Bass, 1980.

Davis, R.H. A Behavioral Change Model with Implications for Faculty Development. Higher Education, 1979, 8, 123-140.

Donnelly, F.A., Ware, J.E., Wolkon, G.H., & Naftulin, D.H. Evaluation of Weekend Seminars for Physicians. Journal of Medical Education, March 1972, 47, 184-187.

Ellis, H.C. The Transfer of Learning. New York: The Macmillan Company, 1965.

Filstead, W.J. Qualitative Methods: A Needed Perspective in Evaluation Research. In T.D. Cook & C.S. Reichardt (Eds.), Qualitative and Quantitative Methods in Evaluation Research. Beverly Hills: Sage Publications, 1979.

Forman, D.C. Evaluation of Training: Present and Future. Educational Technology, October 1980, pp. 48-51.

Gaff, J.G. Toward Faculty Renewal. San Francisco: Jossey-Bass, 1975.

Gaff, J.G. The United States of America--Toward the Improvement of Teaching. In D.C.B. Teather (Ed.), Staff Development in Higher Education: An International Review and Bibliography. New York: Nichols Publishing Company, 1979.

Gaff, J.G., & Morstain, B.R. Evaluating the Outcomes. In J.G. Gaff (Ed.), Institutional Renewal Through the Improvement of Teaching, New Directions for Higher Education, 1978, 6 (4), 77-83.

Giesen, L.A. An Evaluation of Grotelueschen's Model for Evaluating An Adult Education Program. Minneapolis: University of Minnesota, 1979. (ERIC Document Reproduction Service No. ED 201 764)

Glaser, E.M., & Backer, T.E. A Clinical Approach to Program Evaluation. Evaluation: A Forum for Human Service Decision-Makers, Fall 1972, 1 (1), 54-55; 57-58.

Goldstein, I.L. Training: Program Development and Evaluation. Monterey, CA: Brooks/Cole, 1974.

Grenough, J.L., & Dixon, R.W. Using 'Utilization' to Measure Training Results. Training/HRD, February 1982, pp. 40-42.

Grotelueschen, A.D. Program Evaluation. In A.B. Knox & Associates (Eds.), Developing, Administering, and Evaluating Adult Education. San Francisco: Jossey-Bass, 1980.

House, E.R. Assumptions Underlying Evaluation Models. Educational Researcher, March 1978, pp. 4-12.

House, E.R. Evaluating with Validity. Beverly Hills: Sage Publications, 1980.

Howard, G.S., Maxwell, S.E., Wiener, R.L., Boynton, K.S., & Rooney, W.M. Is a Behavioral Measure the Best Estimate of Behavioral Parameters? Perhaps Not. Applied Psychological Measurement, Summer 1980, 4 (3), 293-311.

Howard, G.S., Schmeck, R.R., & Bray, J.H. Internal Invalidity in Studies Employing Self-Report Instruments: A Suggested Remedy. Journal of Educational Measurement, Summer 1979, 16 (2), 129-135.

Hoyt, D.P., & Howard, G.S. The Evaluation of Faculty Development Programs. Research Report Number 39, Manhattan, KS: Kansas State University Office of Educational Research, 1977.

Hull, C.H., & Nie, N.H. SPSS Update 7-9: New Procedures and Facilities for Releases 7-9. New York: McGraw-Hill, 1981.

Hunt, B. Who and What are to be Evaluated? Educational Leadership, January 1978, 35 (4), 260-263.

Joint Committee on Standards for Educational Evaluation. Standards for Evaluations of Educational Programs, Projects, and Materials. New York: McGraw-Hill, 1981.

Joorabchi, B., & Chawhan, A.R. Effects of a Short Educational Planning Workshop on Attitudes of Three Groups of Medical Educators. British Journal of Medical Education, 1975, 9, 38-41.

Kennedy, P.H. Applying Evaluation Standards to Prove Effectiveness of Major Training Program Revisions. Paper presented at the annual meeting of the American Educational Research Association, New York, March 1982.

Kirkpatrick, D.L. Evaluation of Training. In R.L. Craig & L.R. Bittel (Eds.), Training Development Handbook. New York: McGraw-Hill, 1967.

Koen, F.M. A Faculty Educational Development Program and An Evaluation of Its Evaluation. Journal of Medical Education, October 1976, 51, 854-855.

Kryspin, W.J., & Feldhusen, J.F. Developing Classroom Tests: A Guide for Writing and Evaluating Test Items. Minneapolis: Burgess Publishing Company, 1974.

Laird, D. Approaches to Training and Development. Reading, MA: Addison-Wesley, 1978.

Lawson, B.K., & Harvill, L.M. The Evaluation of A Training Program for Improving Residents' Teaching Skills. Journal of Medical Education, December 1980, 55, 1000-1005.

Levinson, J.L., & Menges, R.J. Improving College Teaching: A Critical Review of Research. Occasional Paper Number 13, Evanston, IL: Northwestern University, The Center for the Teaching Professions, 1979.

Levinson-Rose, J.L., & Menges, R.J. Improving College Teaching: A Critical Review of Research. Review of Educational Research, Fall 1981, 51 (3), 403-434.

Littlefield, J.H., Hendricson, B. Kleffner, J.H., & Burns, G. Improving Instructional Skills of Health Science Educators. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Mehrens, W.A., & Lehmann, I.J. Measurement and Evaluation in Education and Psychology (2nd ed.). New York: Holt, Rinehart, & Winston, 1978.

Menges, R.J., & Levinson-Rose, J.L. Why Research Tells Us So Little About Interventions to Improve College Teaching. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Mezoff, B. Pre-Then-Post Testing: A Tool to Improve the Accuracy of Management Training Program Evaluation. NSPI Journal, October 1981, 20 (8), 10-12; 16.

Millman, J. A Checklist Procedure. In N.L. Smith (Ed.), New Techniques for Evaluation, New Perspectives in Evaluation (Volume 2). Beverly Hills: Sage Publications, 1981.

Otto, C.P., & Glaser, R.O. The Management of Training: A Handbook for Training and Development Personnel. Reading, MA: Addison-Wesley, 1970.

Parlett, M., & Hamilton, D. Evaluation as Illumination: A New Approach to the Study of Innovatory Programs. In G.V. Glass (Ed.), Evaluation Studies Review Annual (Volume 1). Beverly Hills: Sage Publications, 1976.

Patton, M.Q. Utilization Focused Evaluation. Beverly Hills: Sage Publications, 1978.

Patton, M.Q. Qualitative Evaluation Methods. Beverly Hills: Sage Publications, 1980.

Posavac, E.J., & Carey, R.G. Program Evaluation: Methods and Case Studies. Englewood Cliffs, NJ: Prentice-Hall, 1980.

Pratt, C.C. Evaluating the Impact of Training on the World of Practice. In P.P. LeBreton, V.E. Bryant, D.L. Zweizig, D.G. Middaugh, A.G. Bryant, & T.C. Corbett (Eds.), The Evaluation of Continuing Education for Professionals: A Systems View. Seattle: University of Washington, 1979.

Ramsey, C.N., Jr., & Hitchcock, M.A. A Statewide Model for Faculty Development in Family Medicine. The Journal of Family Practice, 1980, 11 (3), 421-426.

Reichardt, C.S., & Cook, T.D. Beyond Qualitative Versus Quantitative Methods. In T.D. Cook & C.S. Reichardt (Eds.), Qualitative and Quantitative Methods in Evaluation Research. Beverly Hills: Sage Publications, 1979.

Scriven, M. The Methodology of Evaluation. In R.W. Tyler (Ed.), Perspectives of Curriculum Evaluation. Chicago: Rand McNally, 1967.

Scriven, M. An Introduction to Meta-Evaluation. Educational Product Report, February 1969, 2 (5), 36-38.

Scriven, M. Evaluation Thesaurus (2nd ed.). Inverness, CA: Edgepress, 1980.

Singer, R.N. Motor Learning and Human Performance. New York: The Macmillan Company, 1968.

Smith, N.L. Criticism and Meta-Evaluation. In N.L. Smith (Ed.), New Techniques for Evaluation, New Perspectives in Evaluation (Volume 2). Beverly Hills: Sage Publications, 1981.

Stake, R.E. The Countenance of Educational Evaluation. Teachers College Record, 1967, 68, 523-540.

Steele, S.M. Contemporary Approaches to Program Evaluation and Their Implications for Evaluating Programs for Disadvantaged Adults. Syracuse, NY: ERIC Clearinghouse on Adult Education, 1973.

Stephens, B. A Manual for Clinical Supervision and Facilitative Feedback. Madison, WI: University of Wisconsin Department of Family Medicine and Practice, 1981.

Stevenson, J.F., Longabaugh, R.H., & McNeill, D.N. Metaevaluation in the Human Services. In H.C. Schulberg & J.M. Jerrell (Eds.), The Evaluator and Management, Sage Research Progress in Evaluation (Volume 4). Beverly Hills: Sage Publications, 1979.

Stufflebeam, D.L. Evaluation as Enlightenment for Decision-Making. Columbus, OH: Ohio State University Evaluation Center, 1968.

Stufflebeam, D.L. Toward A Technology for Evaluating Evaluations. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1974.

Stufflebeam, D.L. Metaevaluation: An Overview. Evaluation and the Health Professions, Spring 1978, 1 (1), 17-43.

Stufflebeam, D.L. Metaevaluation: Concepts, Standards, and Uses. In R. Berk (Ed.), Educational Evaluation Methodology: The State of the Art. Baltimore: The Johns Hopkins University Press, 1981.

Taylor, D.B. Eeny, Meeny, Miney, Meaux: Alternative Evaluation Models. The North Central Association Quarterly, Spring 1976, 50 (4), 353-358.

Tyler, R.W. General Statement on Evaluation. Journal of Educational Research, 1942, 36, 492-501.

Tyler, R.W. Basic Principles of Curriculum and Instruction. Chicago: University of Chicago Press, 1949.

Walls, B.E. A Curriculum for Faculty Development in Family Medicine. Unpublished doctoral dissertation, Columbia University Teachers College, 1979.

Warburton, S.W., Frenkel, L., & Snope, F.C. Teaching Physicians to Teach: A Three-Year Report. The Journal of Family Practice, 1979, 9 (4), 649-656.

Wergin, J.F., Mason, E.J., & Munson, P.J. The Practice of Faculty Development: An Experience-Derived Model. Journal of Higher Education, May/June 1976, 47 (3), 289-308.

Wexley, K.N., & Latham, G.P. Developing and Training Human Resources in Organizations. Glenview, IL: Scott, Foresman and Company, 1981.

Worthen, B.R., & Sanders, J.R. Educational Evaluation: Theory and Practice. Worthington, OH: Charles A. Jones Publishing Company, 1973.

LIST OF GENERAL REFERENCES

Adkins, D.C. Test Construction: Development and Interpretation of Achievement Tests (2nd ed.). Columbus, OH: Charles E. Merrill Publishing Company, 1974.

Anderson, S.B., & Ball, S. The Profession and Practice of Program Evaluation. San Francisco: Jossey-Bass, 1978.

Asher, J.W. Educational Research and Evaluation Methods. Boston: Little, Brown and Company, 1976.

Berdie, R.F. Self-Claimed and Tested Knowledge. Educational and Psychological Measurement, 1971, 31 (3), 629-636.

Bergquist, W.H., & Phillips, S.R. A Handbook for Faculty Development (Volume 2). Berkeley, CA: Pacific Soundings Press, 1977.

Bland, C.J. Guidelines for Planning Faculty Development Workshops. The Journal of Family Practice, 1977, 5 (2), 235-241.

Cannell, C.F., & Kahn, R. Interviewing. In G. Lindzey & E. Aronson (Eds.), The Handbook of Social Psychology (2nd ed.). Reading, MA: Addison-Wesley, 1968.

Centra, J.A. Faculty Development in Higher Education. Teachers College Record, Spring 1978, 80 (1), 188-201.

Cook, T.D., & Reichardt, C.S. (Eds.), Qualitative and Quantitative Methods in Evaluation Research. Beverly Hills: Sage Publications, 1979.

Cooley, W.W., & Lohnes, P.R. Evaluation Research in Education. New York: Irvington Publishers, 1976.

Cronbach, L.J. Essentials of Psychological Testing (3rd ed.). New York: Harper & Row, 1970.

Dressel, P.L. Handbook of Academic Evaluation. San Francisco: Jossey-Bass, 1978.

Ebel, R.L. Essentials of Educational Measurement (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall, 1979.

Filstead, W.J. Using Qualitative Methods in Evaluation Research: An Illustrative Bibliography. Evaluation Review, April 1981, 5 (2), 259-268.

Fitz-Gibbon, C.T., & Morris, L.L. How to Design A Program Evaluation. Beverly Hills: Sage Publications, 1978.

Frick, T., & Semmel, M.I. Observer Agreement and Reliabilities of Classroom Observational Measures. Review of Educational Research, Winter 1978, 48 (1), 157-184.

Gorden, R.L. Interviewing: Strategy, Techniques, and Tactics. Homewood, IL: The Dorsey Press, 1975.

Grotelueschen, A.D., Gooler, D.D., & Knox, A.B. Evaluation in Adult Basic Education: How and Why. Danville, IL: The Interstate Printers and Publishers, 1976.

Guba, E.G. Problems in Utilizing the Results of Evaluation. Journal of Research and Development in Education, 1975, 8 (3), 42-54.

Guttentag, M. Models and Methods in Evaluation Research. Journal for the Theory of Social Behavior, April 1971, 1 (1), 75-95.

Hamblin, A.C. Evaluation and Control of Training. London: McGraw-Hill, 1974.

Hays, W.L. Statistics for Psychologists. New York: Holt, Rinehart, & Winston, 1963.

Heilman, J.G. Paradigmatic Choices in Evaluation Methodology. Evaluation Review, October 1980, 4 (5), 693-712.

Hemphill, J.K. The Relationships Between Research and Evaluation Studies. In R.W. Tyler (Ed.), Educational Evaluation: New Roles, New Means, The Sixty-eighth Yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1969.

Herbert, J., & Attridge, C. A Guide for Developers and Users of Observation Systems and Manuals. American Educational Research Journal, Winter 1975, 12 (1), 1-20.

Howard, G.S., Ralph, K.M., Gulanick, N.A., Maxwell, S.E., Nance, D.W., & Gerber, S.K. Internal Invalidity in Pretest-Posttest Self-Report Evaluations and a Re-evaluation of Retrospective Pretests. Applied Psychological Measurement, Winter 1979, 3 (1), 1-23.

Irby, D.M., & Morgan, M.K. Clinical Evaluation: Alternatives for Health Related Educators. Gainesville, FL: University of Florida Center for Allied Health Instructional Personnel, 1975.

Joyce, B.R., & Weil, M. Models of Teaching (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall, 1980.

Lindquist, J. (Ed.). Designing Teaching Improvement Programs. Berkeley, CA: Pacific Soundings Press, 1978.

Loughary, J.W., & Hopson, B. Producing Workshops, Seminars, and Short Courses: A Trainer's Handbook. Chicago: Follett Publishing Company, 1979.

Magnusson, D. Test Theory. Reading, MA: Addison-Wesley, 1966.

Mathis, B.C. What Happened to Research on College Teaching? Journal of Teacher Education, March-April 1980, 31 (2), 17-21.

Mehrens, W.A., & Lehmann, I.J. Standardized Tests in Education. New York: Holt, Rinehart, & Winston, 1973.

Morris, L.L., & Fitz-Gibbon, C.T. How to Measure Achievement. Beverly Hills: Sage Publications, 1978.

Moser, C.A., & Kalton, G. Survey Methods in Social Investigation (2nd ed.). New York: Basic Books, 1972.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., & Bent, D.H. Statistical Package for the Social Sciences (2nd ed.). New York: McGraw-Hill, 1975.

Popham, W.J. Educational Evaluation. Englewood Cliffs, NJ: Prentice-Hall, 1975.

Roid, G.H., & Haladyna, T.M. A Technology for Test-Item Writing. New York: Academic Press, 1982.

Rossi, P.H., Freeman, H.E., & Wright, S.R. Evaluation: A Systematic Approach. Beverly Hills: Sage Publications, 1979.

Saint Louis University. A Model for Assessing the Quality of Non-Traditional Programs in Higher Education. St. Louis: Saint Louis University, 1979.

Scriven, M. Evaluation Bias and Its Control. In G.V. Glass (Ed.), Evaluation Studies Review Annual (Volume 1). Beverly Hills: Sage Publications, 1976.

Shenson, H.L. How to Create and Market A Successful Seminar or Workshop. Washington, DC: Bermont Books, 1981.

Smith, J.M. Interviewing in Market and Social Research. London: Routledge & Kegan Paul, 1972.

Smith, M.E., O'Callaghan, M., Corbett, A.J., Morley, B., & Kamradt, I.L. An Example of Meta-Evaluation from Industry. Improving Human Performance Quarterly, 1977, 5 (3-4), 168-182.

Staropoli, C.J., & Waltz, C.F. Developing and Evaluating Educational Programs for Health Care Providers. Philadelphia: F.A. Davis, 1978.

Suchman, E.A. Evaluative Research. New York: Russell Sage Foundation, 1967.

Thornton, G.C., III. Psychometric Properties of Self-Appraisals of Job Performance. Personnel Psychology, Summer 1980, 33 (2), 263-271.

Weiss, C.H. Evaluation Research: Methods for Assessing Program Effectiveness. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Westberg, J., & Jason, H. The Enhancement of Teaching Skills in U.S. Medical Schools: An Overview and Some Recommendations. Medical Teacher, 1981, 3 (3), 100-104.

APPENDICES

APPENDIX A

BACKGROUND INFORMATION ON THE FAMILY MEDICINE FACULTY DEVELOPMENT PROGRAM

FAMILY MEDICINE FACULTY DEVELOPMENT FELLOWSHIP PROGRAM

Michigan State University
September 8-18, 1981

MONDAY 7	TUESDAY 8	WEDNESDAY 9	THURSDAY 10	FRIDAY 11
LABOR DAY HOLIDAY	8:30-11:30 a.m. Welcome and Orientation (Anderson & Henry) 1:30-4:30 p.m. Elements of Group Development (Casbergue)	8:30-11:30 a.m. Principles of Learning and Motivation (Yelon) 1:30-4:30 p.m. Principles of Learning and Motivation (Yelon)	8:30-11:30 a.m. Clinical Teaching Technique (Farquhar & Casbergue) 1:30-4:30 p.m. Curriculum Development in Family Medicine (Hough & McKeag)	8:30-10:00 a.m. Issues in Family Medicine (Gerard & Greenman) 10:00-11:30 a.m. Producing Audiovisual Materials (Andrews, Schulz & Wyckoff) 1:30-4:00 p.m. Producing Audiovisual Materials 4:00-4:30 p.m. End-of-Week Evaluation (Anderson)

MONDAY 14	TUESDAY 15	WEDNESDAY 16	THURSDAY 17	FRIDAY 18
8:30-11:30 a.m. Teaching Psychomotor Skills (Yelon) 1:30-4:30 p.m. Psychomotor Skill Teaching Practice (Yelon)	8:30-11:30 a.m. Presentation Skills (Anderson) 1:30-4:30 p.m. Role of Clinical Supervision (Casbergue & Farquhar)	8:30-11:30 a.m. Perspectives in Learning (Henry & Rothert) 1:30-4:30 p.m. Constructive Feedback in Clinical Education (Casbergue & Farquhar)	8:30-11:30 a.m. Clinical Teaching Simulation (Casbergue & Farquhar) 1:30-4:30 p.m. PRACTICUM	8:30-11:30 a.m. Asking and Answering Student Questions (Anderson) 1:30-4:00 p.m. Practice Teaching (Anderson & Henry) 4:00-4:30 p.m. End-of-Week Evaluation (Anderson)

Who We Are....

The Family Medicine Faculty Development Program at Michigan State University began operation in July, 1978. The program is conducted by the Office of Medical Education Research and Development (OMERAD), in association with the departments of Family Practice (College of Human Medicine) and Family Medicine (College of Osteopathic Medicine) at Michigan State University. This program is supported by a grant from the Bureau of Health Manpower, Public Health Service. Michigan State University's program addresses two major objectives. They are:

1. to identify and train new physician teaching faculty for both allopathic and osteopathic family medicine training programs, and
2. to assist existing family medicine faculty in developing and/or refining their pedagogical skills.

These objectives are being met through three distinct yet coordinated program components. These components are: 1) a series of teaching skills workshops; 2) a teaching fellowship; and 3) a continuing professional development program.

Teaching Skills Workshops

The teaching skills workshops are being offered to M.D. residents, D.O. interns, preceptors, and other part-time physician faculty who have informal teaching responsibilities in family medicine. The purpose of these workshops is to mobilize interest in teaching as a career and to improve the teaching skills of these physicians. The workshops are no longer than one day and focus on specific instructional planning, teaching, and evaluation skills. Each year, a total of eight workshops will be conducted for M.D. and D.O. physicians in the Michigan area.

Teaching Fellowship

A teaching fellowship is being offered to M.D. and D.O. physicians who have completed or are about to complete a family medicine residency program and to family medicine physicians who are just beginning their teaching career. The fellowship begins in September, and fellows spend one- and two-week sessions at Michigan State University throughout the remainder of the academic year. The fellowship is the equivalent of a three-month traineeship.

The goal of the fellowship program is to provide these new faculty members with a proven base of skills in teaching, evaluation, and the management of instruction. Fellows participate in a series of workshops, seminars, and practice teaching situations in real and simulated clinical, lecture, and small group settings. A portion of the fellowship program is conducted at the fellow's home institution. Here, fellows complete a variety of structured assignments under the supervision of a project faculty member. A stipend is available to help fellows defer the costs of participating in the program.

Continuing Professional Development Program

The continuing professional development program will be offered to existing M.D. and D.O. physician faculty with regular teaching responsibilities in family medicine training programs. Beginning in 1982-83, the purpose of this program will be to reduce the rate at which full-time faculty members leave the teaching of family medicine and to provide a forum for the continuing professional development of faculty who cannot participate in the three-month traineeships. The program will include a series of interactive seminars that will allow full-time faculty members to meet with faculty from different institutions to systematically develop solutions to chronic problems in the teaching of family medicine. The seminars will meet approximately ten times during a year in various residency program and family medicine departmental settings.

Faculty Development Workshop Materials

In addition to providing formal training programs, the Family Medicine Faculty Development Program at Michigan State University has developed a series of eight self-standing mediated faculty development workshops. The purpose of these workshops is to assist family medicine departments and residency programs in conducting their own faculty development programs. Each workshop package contains all the print and audiovisual materials necessary to conduct the workshop. A detailed administrator's guide explains all steps necessary for planning, conducting, and evaluating each workshop.

For additional information about the activities of Michigan State University's Family Medicine Faculty Development Program, please contact:

Dr. William A. Anderson
Office of Medical Education
Research and Development
A-209 East Fee Hall
Michigan State University
East Lansing, Michigan 48824
Phone: (517) 353-9656

APPENDIX B

END-OF-WEEK EVALUATION FORMS

FAMILY MEDICINE FACULTY DEVELOPMENT PROGRAM

End-of-Week Evaluation
Week 1
September 11, 1981

PART I

Please indicate your overall reactions to this past week's sessions by checking the appropriate box. If you have a specific suggestion about how a change should be made, write that suggestion in the appropriate space or at the end of this instrument.

Aspect of Program	Keep the Same	Increase	Decrease	Specific Suggestion
1. Amount of Reading				
2. Comfort of Room				
3. Length of Workshops				
4. Relevance of Information				
5. Amount of Participation				
6. Amount of Practice				
7. Number of Examples				
8. Level of Information Compared to My Amount of Knowledge				

PART II

Please respond to the following statements using the KEY given below.

KEY: SA means you strongly agree with the statement, A means you agree, U means you are uncertain, D means you disagree, and SD means you strongly disagree.

1. I found the Tuesday orientation session very helpful.

SA A U D SD

2. The concepts presented in the small group process workshop were helpful.

SA A U D SD

3. I believe I can effectively use the principles of learning and motivation in my own teaching.

SA A U D SD

4. The clinical teaching technique session was helpful for understanding my own teaching style and preferences.

SA A U D SD

5. The session on curriculum development in sports medicine was useful.

SA A U D SD

6. I have a better understanding of the similarities and differences between allopathic and osteopathic family medicine.

SA A U D SD

7. I can use some of the ideas and skills from the audiovisual workshop.

SA A U D SD

PART III

Please write your responses to the following questions in the space provided.

1. What was the most helpful presentation or discussion during this past week?

2. What presentation or discussion during this past week was not relevant to your needs?

3. What part of the program gave you the most difficulty?

4. What can we do to help you learn during the program?

5. What other suggestions do you have for improving the program?

6. What is your overall reaction to this week's program?

ADDITIONAL COMMENTS

FAMILY MEDICINE FACULTY DEVELOPMENT PROGRAM

End-of-Week Evaluation
Week 2
September 18, 1981

PART I

Please indicate your overall reactions to this past week's sessions by checking the appropriate box. If you have a specific suggestion about how a change should be made, write that suggestion in the appropriate space or at the end of this instrument.

Aspect of Program	Keep the Same	Increase	Decrease	Specific Suggestion
1. Amount of Reading				
2. Comfort of Room				
3. Length of Workshops				
4. Relevance of Information				
5. Amount of Participation				
6. Amount of Practice				
7. Number of Examples				
8. Level of Information Compared to My Amount of Knowledge				

PART II

Please respond to the following statements using the KEY given below.

KEY: SA means you strongly agree with the statement, A means you agree, U means you are uncertain, D means you disagree, and SD means you strongly disagree.

1. As a result of the psychomotor teaching skills session, I am better prepared to teach these types of skills.

SA A U D SD

2. I will use the approach presented for the teaching of psychomotor skills.

SA A U D SD

3. The session on presentation skills will help me improve my own presentations.

SA A U D SD

4. I feel more skilled as a clinical supervisor.

SA A U D SD

5. I will use the ideas presented in the constructive feedback session.

SA A U D SD

6. I am more aware of my own thinking as a physician as a result of the discussion on perspectives in learning.

SA A U D SD

7. I found the clinical teaching practice teaching sessions (videotape) helpful.

SA A U D SD

8. I have a better idea of how to ask and answer student questions.

SA A U D SD

9. The "practicum" session (Thursday afternoon) should be continued.

SA	A	U	D	SD

10. I found the practice teaching assignment a valuable learning experience.

SA	A	U	D	SD

PART III

Please write your responses to the following questions in the space provided.

1. What was the most helpful presentation or discussion during the past week?

2. What presentation or discussion during this past week was not relevant to your needs?

3. What information or topic(s) on teaching was missing from this two-week session?

4. What suggestions do you have for improving this two-week session?

5. What is your overall reaction to this two-week session?

6. What research and evaluation topics would you like to see addressed in the January session?

ADDITIONAL COMMENTS

APPENDIX C

COGNITIVE PRETEST

Family Medicine Faculty Development Program
Michigan State University

September 8, 1981

PRETEST

Instructions:

This pretest is part of a program evaluation of the Family Medicine Faculty Development Fellowship Program. Answer each item as completely as possible. Consideration will be given for partially correct responses. Thank you for your cooperation.

Name _____

Small Group Development

1. List two components or dimensions of group development that a group leader should be aware of in order to monitor and influence the development of a small group.

2. List three characteristics of effective group functioning.

3. State three advantages of appropriately using a group discussion format.

4. What types of actions might a group leader take when there is conflict (e.g., strongly opposing views, arguing, etc.) during a group discussion/meeting?

Principles of Learning and Motivation

5. What techniques can be used to make learning meaningful to your students?

6. Describe what techniques you could use during a lesson to stimulate your students' attention.

7. What is the most appropriate use of modeling in instruction?

8. Describe how you could go about establishing and maintaining open communication between you and your students.

9. What must be considered in determining which medium and strategy are appropriate for your presentation?

Clinical Teaching Technique

10. Describe three roles that a clinical instructor might use in teaching medical students or residents.

11. What three topics should be discussed by the instructor immediately before observing an initial student/patient contact?

12. What are four characteristics of effective feedback?

13. How might a lack of feedback negatively effect the learner? (Describe three)

14. What factors might inhibit the process of giving feedback:

    a. from the teacher to the learner?

    b. from the learner to the teacher on his/her role as a teacher?

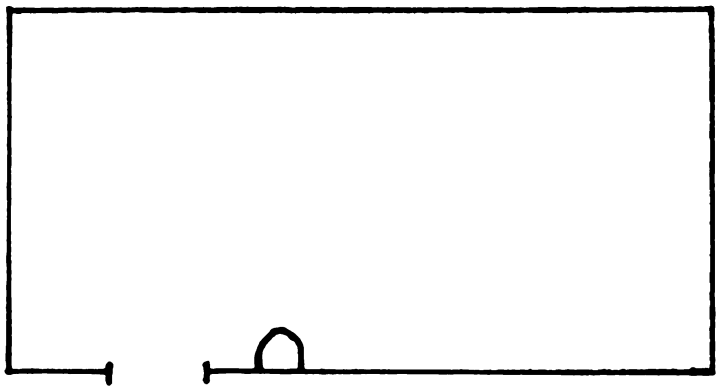
15. Given a hypothetical six-week period during which you will be working closely with a student, which teaching technique will you probably use the most during weeks 1 and 6?

16. What are three goals of clinical supervision?

17. List three strategies that a clinical teacher might use in dealing with an anxious patient.

18. Physically arrange the examining room and position the following persons to provide an effective history gathering and/or examination opportunity for the resident. Use symbols to show desired position; point indicates direction facing or sitting.

    a. Patient
    b. Resident
    c. Clinical teacher
    d. Desk
    e. Examination table



Producing Audiovisual Materials

19. List four ways television can be used in undergraduate medical education.

20. List four steps in the process of selecting the appropriate media format for a presentation.

21. State a "rule of thumb" for the maximum amount of printed material to be used on a projected visual image.

22. List two advantages and two disadvantages of 35mm slides, overhead transparencies, and television in an instructional mode.

Teaching Psychomotor Skills

Situation: You have been assigned to teach a first-year medical student how to use and read a sphygmomanometer.

23. Why is the use of objectives important in introducing a student to the proper use of the instrument?

24. How would you determine prerequisite or entry skills before teaching the student how to use this instrument?

25. What steps would you include when you introduce and demonstrate any new psychomotor skill to a student?

26. What should you tell your student prior to a demonstration of the use of the sphygmomanometer? Why?

27. What should you have your student do immediately following the demonstration?

28. Describe how and when you would provide feedback to a student practicing the use of the sphygmomanometer.

29. Based on your experience as a family medicine physician, cite two different psychomotor skills that you could teach to first-year medical students.

Presentation Skills

Situation: You have been asked to give a lecture presentation to a group of businessmen on the topic of hypertension.

30. What should you try to accomplish in the introduction of your presentation?

31. What techniques could you use to get the audience actively involved during the lecture?

32. In what ways could you determine how much of the lecture has been understood by the audience?

33. How would you determine the audience's personal reactions and attitudes toward your lecture? Write two specific questionnaire items you would use to elicit this information.

34. Describe a situation in which a lecture would be an appropriate instructional method and explain why.

Perspectives in Learning

35. What are three different types of learning a clinical faculty member would encounter?

36. Distinguish the cognitive view of learning from the behavioral view.

37. How might you draw an incorrect clinical conclusion by using an "availability" heuristic?

38. What are the three factors affecting the predictive value of a laboratory test in a clinical setting?

39. Describe the overall goal of applying decision analysis in clinical practice, i.e., why might decision analysis be useful to physicians?

40. What does it mean to say that physicians work with probabilistic information or make judgments under uncertainty? Include at least one example in your response.

APPENDIX D

COGNITIVE TEST RATING SCALE

Cognitive Test Rating Scale

0 - nothing present or completely wrong

1 - minimal answer, but something there is correct

2 - more than one component present or correct

3 - nearly all components present or correct or all correct or present

APPENDIX E

VIDEOTAPE RATING SCALE

Videotape Rating Scale

FELLOW_____

SESSION: JAN___ MAY___

DATE RATED_____

SCALE:
0 - Not done
1 - Done poorly
2 -
3 - Done moderately well
4 -
5 - Done very well

Using the above scale, circle the appropriate number for each of the following items.

In this videtaped presentation, the fellow:

1.	Introduced the topic of the presentation.	0	1	2	3	4	5
2.	Related the presentation to the audience members' past, present, or future.	0	1	2	3	4	5
3.	Provided necessary aids to organize the presentation.	0	1	2	3	4	5
4.	Delivered, rather than read, the presentation.	0	1	2	3	4	5
5.	Presented information in an organized, logical manner.	0	1	2	3	4	5
6.	Used examples or illustrations.	0	1	2	3	4	5
7.	Showed interest in the topic and enthusiasm in presenting it.	0	1	2	3	4	5
8.	Summarized main points or ideas of the presentation.	0	1	2	3	4	5
9.	Maintained eye contact with members of the audience.	0	1	2	3	4	5
10.	Maintained good posture throughout the presentation.	0	1	2	3	4	5
11.	Varied the rate and pace of the presentation.	0	1	2	3	4	5
12.	Used appropriate gestures during the presentation.	0	1	2	3	4	5

13. Spoke clearly and audibly.	0	1	2	3	4	5
14. Provided smooth transitions between main ideas or points.	0	1	2	3	4	5
15. Solicited ideas or questions from members of the audience.	0	1	2	3	4	5
16. Responded to ideas or questions from members of the audience.	0	1	2	3	4	5

COMMENTS/NOTES:

APPENDIX F

INTERVIEW PROTOCOLS

FELLOW INTERVIEW QUESTIONNAIRE

NAME_____

DATE_____

QUESTION #1

The first question deals with your previous knowledge of the content of the sessions that were conducted in September. I will read you the name of each topic and then will ask you to respond either "Yes" or "No" to the question that I will ask you about each topic. If you respond "yes" then I will ask you to rate your expertise in that topic PRIOR to the beginning of the September program. In making that rating you should keep in mind that a scale from 1 to 5 will be used, with 1 low, 3 medium, and 5 high. Any questions? (PAUSE) Okay, here goes.

Before the September program did you have a background in or any previous experience with:

elements of group development	Y___	N___	1	2	3	4	5
clinical teaching technique	Y___	N___	1	2	3	4	5
role of clinical supervision	Y___	N___	1	2	3	4	5
constructive feedback in clinical education	Y___	N___	1	2	3	4	5
principles of learning and motivation	Y___	N___	1	2	3	4	5
teaching psychomotor skills	Y___	N___	1	2	3	4	5
producing audiovisual materials	Y___	N___	1	2	3	4	5
presentation skills	Y___	N___	1	2	3	4	5
asking and answering student questions	Y___	N___	1	2	3	4	5
perspectives in learning	Y___	N___	1	2	3	4	5

(If yes for any of the above) How would you rate your expertise in this topic prior to the September program on a scale from 1 to 5 with 1 low and 5 high?

QUESTION #2

For the second question I would like to ask you if you have undertaken any additional study, such as reading, workshop attendance, CME activities, or other methods of study or learning, in any of the topics of the September program since that program ended. As in the previous question I will read you the question and the topic and then ask you to respond either "Yes" or "No." I will also ask you to rate your current expertise in each of the topics. Again the scale will be from 1 to 5, with 1 low and 5 high. Any questions? (PAUSE) Okay, here goes.

Since the September program ended have you undertaken any additional study in:

elements of group development	Y___	N___	1	2	3	4	5#
clinical teaching technique	Y___	N___	1	2	3	4	5
role of clinical supervision	Y___	N___	1	2	3	4	5
constructive feedback in clinical education	Y___	N___	1	2	3	4	5
principles of learning and motivation	Y___	N___	1	2	3	4	5
teaching psychomotor skills	Y___	N___	1	2	3	4	5
producing audiovisual materials	Y___	N___	1	2	3	4	5
presentation skills	Y___	N___	1	2	3	4	5
asking and answering student questions	Y___	N___	1	2	3	4	5
perspectives in learning	Y___	N___	1	2	3	4	5

#How would you rate your expertise in this topic at this moment?

QUESTION #3

Have you used any of your notes or handouts from the September program since that program ended?

Y___ N___

(If yes) For which topics have you used your notes or handout materials?

elements of group development	Y___ N___	How often?	once	2-3	4-5	5+
clinical teaching technique	Y___ N___	How often?	once	2-3	4-5	5+
role of clinical supervision	Y___ N___	How often?	once	2-3	4-5	5+
constructive feedback in clinical education	Y___ N___	How often?	once	2-3	4-5	5+
principles of learning and motivation	Y___ N___	How often?	once	2-3	4-5	5+
teaching psychomotor skills	Y___ N___	How often?	once	2-3	4-5	5+
producing audiovisual materials	Y___ N___	How often?	once	2-3	4-5	5+
presentation skills	Y___ N___	How often?	once	2-3	4-5	5+
asking and answering student questions	Y___ N___	How often?	once	2-3	4-5	5+
perspectives in learning	Y___ N___	How often?	once	2-3	4-5	5+

QUESTION #4

Have you shared your new knowledge and skills that you learned during the September program with your colleagues or other people in your organization or community?

Y___ N___

(If yes) Which of the following categories best describe how you shared your new knowledge and/or skills? You may choose more than one category if it is appropriate to your situation.

The categories are:

_____formal presentation
_____individual consultation
_____informal conversation(s)
_____written communication(s)
_____other (please specify)

QUESTION #5

In the six months since the end of the September program, have you had an opportunity to use any of the knowledge or skills that you learned during those two weeks?

Y___ N___

(If yes) Please describe what specific knowledge or skills you have been able to use.

Now please describe how you were able to use this specific knowledge or skills.

QUESTION #6

In the next six months, do you expect to have an opportunity to use any of the knowledge or skills you learned during the September program?

Y___ N___

(If yes) Please describe what specific knowledge or skills you expect to be able to use.

Now please describe how you expect to be able to use this knowledge or skill.

The next series of questions is concerned with the exercises or simulations that you participated in at MSU that were videotaped for you to review at a later time.

QUESTION #7

Did you review the videotape in which you were placed in the role of a clinical teacher supervising a first-year resident?

Y___ N___

(If yes) On a scale from 1 to 5, with 1 low and 5 high, how would you rate your overall perfomance as a clinical teacher in that videotape?

1 2 3 4 5

If you were to go through that same clinical teaching simulation tomorrow, how would you rate your expected performance? Again, use a scale from 1 to 5.

1 2 3 4 5

QUESTION #8

Did you review the videotape of your presentation assignment? If you remember, that was the one on the last Friday of the September seesion where you were asked to teach something to someone.

Y___ N___

(If yes) On a scale from 1 to 5, how would you rate your overall performance in that presentation assignment?

1 2 3 4 5

If you were given the same assignment, to teach something to someone for twenty minutes, and you had to do it tomorrow, how would you rate your expected performance?

1 2 3 4 5

QUESTION #9

Did you review the videotape of the research and evaluation project presentation that you gave in January?

Y___ N___

(If yes) On a scale from 1 to 5, how would you rate your overall performance in that presentation? Note that the focus is on your presentation and the associated skills, not on the content of the research or evaluation project that your presented.

1 2 3 4 5

If you have to give a similar presentation tomorrow, how would you rate your expected performance?

1 2 3 4 5

QUESTION #10

Has your participation in the September program changed your role or function in your organization? For example, have your tried some new teaching techniques or have you significantly changed any of your daily activities?

Y___ N___

(If yes) Please describe how your role or function has changed.

QUESTION #11

Has your perception of teaching as a career changed since the completion of the September program?

Y___ N___

(If yes) Please describe how your perception of teaching as a career has changed.

QUESTION #12

Do you feel that the September program has helped you to become a better teacher?

Y___ N___

(If yes) Please describe how the program has helped you become a better teacher.

(If no) Please describe why the progam has not helped you become a better teacher.

QUESTION #13

If a friend or acquaintance of yours was interested in becoming a faculty member in family medicine or wanted to become a better teacher of family medicine, would you recommend the September program to him/her?

Y___ N___

(If yes) Why would you recommend the September program to someone interested in becoming a faculty member in family medicine?

(If no) Why wouldn't you recommend the September program to someone interested in becoming a faculty member in family medicine?

QUESTION #14

Is there anything else that has happened to you as a result of the September program that has not been covered by these questions?

Y___ N___

(If yes) Please explain or describe.

QUESTION #15

Do you have any additional comments or concerns that you wish to express at this time?

Y___ N___

(If yes) Please make them at this time.

CLOSING REMARKS:

That completes the questions that I have for you at this time. As I mentioned at the beginning of the interview, your responses and comments will remain confidential and no names will be used in the final evaluation report. Since many of you have expressed an interest in hearing the results of the evaluation, I will be presenting the results sometime during the May session. Thank you very much for your time and cooperation throughout both this interview and the times when you were taking the written test. Without your cooperation, a quality evaluation of this program would not be possible. Again, thank you very much for taking the time to talk with me at this time.

SUPERVISOR INTERVIEW QUESTIONNAIRE

NAME_____

DATE_____

QUESTION #1

Once Dr. _____ learned about the FMFD Program, did you encourage him/her to participate in the program?

Y___ N___

(If yes) Why?

(If no) Why not?

QUESTION #2

Has Dr. _____ shared any of the information or new knowledge or skills that he/she learned about teaching during the September program with you or other members of your organization?

Y___ N___

(If yes) Which of the following method or methods best describe how he/she shared this information?

____formal presentation
____individual consultation
____informal conversations
____written communication
____other (please specify)

QUESTION #3

Do you know if Dr. _____ has been able to use any of the new knowledge or skills related to teaching that he/she learned in September at MSU?

Y___ N___

(If yes) Please describe the types of knowledge and/or skills that Dr. _____ has been able to use, and the types of situations that they have been used in.

QUESTION #4

Have you observed Dr. _____ doing any teaching since late September? This could include activities such as one-on-one clinical teaching or supervision, small group discussion teaching, or formal lectures or presentations.

Y___ N___

(If yes) How often have you observed Dr. _____ doing some teaching since late September?

___once
___2 to 3 times
___4 to 5 times
___more than 5 times

QUESTION #5

Do you feel able to judge whether or not Dr. _____'s teaching behavior has changed since late September?

Y___ N___

(If yes) How has Dr. _____ 's teacher behavior changed since September?

What do you think has caused the change in Dr. _____'s teaching behavior?

QUESTION #6

Have you noticed any change in Dr. _____'s role or function in your organization since the end of the September program? For example, has he/she become active in new areas of your program or has he/she taken on new responsibilities?

Y___ N___

(If yes) Please describe this change in Dr. _____'s role or function.

QUESTION #7

Has your program benefited in any way by Dr. _____'s participation in the FMFD Program?

Y___ N___

(If yes) Please describe how your program has benefited.

(If no) Please explain why you do not believe that your program has benefited.

QUESTION #8

Would you encourage another resident (or faculty member) from your program to participate in the fellowship program in the future?

Y___ N___

(If yes) Why?

(If no) Why not?

QUESTION #9

Do you have any additional comments about either Dr. _____'s teaching behavior or skills or about the FMFD Program that you would like to make at this time?

Y___ N___

QUESTION #10

Would you like to receive a copy of the final evaluation report on the FMFD Program?

Y___ N___

(If yes) I will arrange for you to receive a final copy of the evaluation report.

QUESTION #11

Do you have any other comments or concerns that you would wish to express at this time?

Y___ N___

CLOSING

That completes the interview. Thank you for your time and cooperation. Hopefully the results of this program evaluation can be used to improve the FMFD Program so that people like yourself will continue to send prospective teachers of family medicine to participate in the program. Thanks again for your time and comments.

PROGRAM DIRECTOR INTERVIEW QUESTIONNAIRE

1. Did you have any concerns about the September session of the FMFD Program before it started? For example, were there any new segments, new faculty, resource constraints, or other possible problems?

2. Did you have any concerns about the participants prior to the September session? For example, were you worried about the size of the group, the MD-DO mix, the resident-faculty mix?

3. How did you feel after the completion of the September session? Were you satisfied with the individual segments, faculty, participants, or any other aspects of the session?

4. Based on your first impressions from reading their applications, talking to their supervisors, meeting them for the first time, or using any other information you had, who would you have picked as the fellows most likely to do well in the activities of the September session? Least likely to do well?

5. After observing the fellows during the two weeks in September, who appeared to have mastered the skills and techniques of that session (or perhaps had arrived on the scene with them already)? Who had made the most improvement over the two-week period of time? Had anyone slid back, regressed?

6. When the fellows came back in January and gave their proposed project presentations, who appeared to be the most skillful and effective presenters? Who had made the most improvement since September? Who had regressed or remained the same?

7. When they returned in May and gave their final project presentations, who appeared the most skillful and effective? Who had made the biggest improvement since January? Since September? Who were the biggest suprises, either positive or negative, to you over time from September to May?

8. Both of you worked closely with the fellows in preparing various presentations and in conducting their major projects. Which fellows showed during those contacts that they had a good command of the terminology, concepts, techniques, and skills covered during the September session?

9. Did any of the fellows do any follow-up work with you related to the topics of the September session? Did you supply any of them with any additional handouts, references, or any other information related to the September session?

10. Have you noticed or learned of any unintended or unplanned outcomes among the fellows as a result of the September session?

11. How would you compare the overall quality of the presentation skills of this group of fellows (based on the major project presentations) with those of previous groups of fellows? How would you explain this?

12. Do you have any additional comments to make concerning the fellows related to the activities of the September session?

APPENDIX G

FINAL DEBRIEFING QUESTIONNAIRE

FAMILY MEDICINE FACULTY DEVELOPMENT PROGRAM
1981-82 FINAL DEBRIEFING

PART I: Written Responses

Please write your responses to the following questions in the space provided.

1. What is your overall evaluation of the program?

2. What was missing most from the program?

3. What comments do you have about the administration of the program?

4. How would you rate your contribution to the program?

5. What would better prepare fellows for the program?

6. What comments do you have about the evaluation of the program? (pre/posttest, telephone interviews, end-of-week evaluations)

7. Comments

PART II: Discussion Topics

1. Major Projects
2. September Session: "Teaching and Learning"
3. January Session: "Research and Evaluation"
4. March Session: "Issues in Family Medicine"
5. May Session: "Administrative Skills"

APPENDIX H

METAEVALUATION PROCEDURE: PROGRAM DIRECTOR INTERVIEW

METAEVALUATION PROCEDURE:
PROGRAM DIRECTOR INTERVIEW

RESEARCH QUESTION: Was the evaluation framework practical in its use of resources?

SPECIFIC QUESTIONS:

1. Did the evaluation procedures produce information of sufficient value to justify the resources expended?

2. Were the evaluation procedures administered so that program disruption was kept to a minimum?

3. Did the use of multiple instruments appear to yield results that justified the extra time and effort involved in their development and administration?

RESEARCH QUESTION: Was the evaluation framework valuable in providing information to you as decision makers?

4. Did it provide information that answered specific questions that you had about the program?

5. Was the information that you received complete and comprehensive? Was there anything left out that you would like to have known?

6. How could the evaluation have been changed to provide more useful information?

RESEARCH QUESTION: Were the methods and instruments used within the evaluation framework technically adequate?

7. Were the sources of information described in enough detail for you to assess the validity of the information they provided?

8. Were the information-gathering instruments and procedures described in enough detail for you to assess the validity of the results they produced?

9. Were the information-gathering instruments and procedures described in enough detail for you to assess the reliability of the results they produced?

10. Was there evidence that the data were collected and analyzed systematically?

11. Did it appear that the quantitative data were appropriately and systematically analyzed?

12. Did it appear that the qualitative data were appropriately and systematically analyzed?

13. Were the conclusions presented in the evaluation report supported by the data?

RESEARCH QUESTION: Were the methods and instruments used within the evaluation framework ethical in dealing with people and organizations?

14. Was the evaluation report open, direct, and honest in its disclosure of pertinent findings, including the limitations of the evaluation?

15. Was the evaluation designed and conducted so that the rights and welfare of the human subjects were respected and protected?

ADDITIONAL QUESTIONS

16. What would you do if you were to do it again?

17. What was not done that should be done?

18. Which evaluation source provided the information of most value to you?

19. Which evaluation method provided the information of most value to you?

20. Which type of data would you rely on?

21. Which type of data would you least rely on?

22. What was the overall strength of the evaluation?

23. What are you doing this time?

APPENDIX I

EVALUATION REPORT: INTRODUCTION

EVALUATION OF THE SEPTEMBER SESSION OF THE FAMILY MEDICINE FACULTY DEVELOPMENT PROGRAM

ACADEMIC YEAR, 1981 - 82

EVALUATION REPORT

PREPARED BY:

Kent J. Sheets, Ph.D. (Cand.)

September 12, 1982

INTRODUCTION

The Family Medicine Faculty Development Program (FMFDP) conducted by the Office of Medical Education Research and Development (OMERAD) at Michigan State University (MSU) is supported by a grant from the Bureau of Health Manpower, Public Health Service. The two major objectives of this program are to identify and train new physician teaching faculty for family medicine training programs and to help current family medicine faculty develop and/or refine their teaching skills. One component of this program is a three-month teaching fellowship offered to allopathic (M.D.) and osteopathic (D.O.) physicians who have completed or are near completion of a family medicine residency program and to family medicine physicians with one year or less of academic teaching experience. It is a two-week session of this fellowship that is the subject of this evaluation report.

The goal of the fellowship is to provide the fellows with a foundation of skills in teaching, evaluation, and the management of instruction. The fellowship begins in September, and participants spend one- and two-week sessions at MSU throughout the remainder of the academic year. During these sessions at MSU the fellows participate in a series of workshops, seminars, and practice teaching situations conducted by nationally known medical educators. A stipend is available to help fellows cover the costs of participating in the fellowship.

The two-week session of the fellowship that was evaluated was conducted in September 1981. This session presented workshops and activities concerned specifically with techniques and principles related to teaching and learning in medical schools and residency training programs. A copy of the schedule for the September 1981 session is included in the

appendices.

The FMFDP has been in operation since July 1978 and has been successful in meeting its goal of increasing the number of family physicians in academic positions, but the program directors have little empirical evidence demonstrating that the program has had an impact on the knowledge, skills, and performance of the participants. Therefore, the evaluation described in this report was conducted by the author in an attempt to determine whether the September session had an impact on the participants and/or their organizations. The evaluator utilized evaluation procedures already in use by the FMFDP staff and also developed some new evaluation instruments in order to gather different types of evaluation data from a variety of sources.

One objective of this evaluation was to provide information to the FMFDP Directors to help them improve and plan future offerings of the September session. A second objective was to determine whether there was any kind of evidence that the September session benefited the participants and/or their home institutions. While the emphasis of the evaluation focused on meeting these two objectives, it was also intended to explore alternative evaluation procedures that might be more effective in gathering information useful to the FMFDP Directors and also to look for any unintended or unexpected outcomes.

In the remainder of the evaluation report, the evaluation procedures used to gather the data presented in this report are briefly described. Examples of the evaluation instruments used to gather the data are provided in the appendices. The evaluation questions that were formulated to guide this study are presented with summaries of the results that correspond to each question or group of questions. Complete data sets

appear in the appendices. In closing, a summary is provided outlining overall results of the evaluation. Recommendations and other comments are also presented.

APPENDIX J

FIELDTEST DATA: END-OF-WEEK EVALUATIONS

FAMILY MEDICINE FACULTY DEVELOPMENT PROGRAM

End-of-Week Evaluation
Week 1
September 8-11, 1981

SUMMARY

PART I

	Aspect of Program	Keep the Same	Increase	Decrease	Specific Suggestion
1.	Amount of Reading	15	1	2	Hard to keep track of what all the different handouts relate to. Maybe color code or # or index; More references, but continue to show priorities among references
2.	Comfort of Room	8	10		Need windows in room; Could be better; I prefer the basement, but we should change rooms occasionally; Too congested area: Room cramped for the number of people; Could have more room; Subdue lighting slightly;

	Aspect of Program	Keep the Same	Increase	Decrease	Specific Suggestion
					Room size was a little small (216);
					It's really fine, but windows are pleasanter if possible
3.	Length of Workshops	15		3	More regular breaks (10 minute);
					Decrease the a.m. audiovisual session. Keep the movie, but a lot of the a.m. audiovisual session was covered in the p.m.;
					Sports medicine lengthy;
					Some seemed a little long, others were okay.
4.	Relevance of Information	13	5		Decrease Thursday p.m. and Friday a.m. sessions;
					Could take more information time;
					Keep the same except sports medicine;
					Information seems vague, hard to apply;
					More details on osteopathic/allopathic, improve curriculum development;

	Aspect of Program	Keep the Same	Increase	Decrease	Specific Suggestion
					Orient lecture topics as you go--not for the whole fellowship at one time; Sports medicine--important issues raised, but not all questions thoroughly answered or prepared for, i.e., evaluation, personal/family needs of the busy practitioner in evaluation
5.	Amount of Participation	15	2	1	A little too much
6.	Amount of Practice	15	3		Need ways to apply information--everything sounds good, but I'm left wondering how to use it; Increase for audio-visual; Good for this stage; Practice (mini sessions) important, but not always clear, maybe do a "role play" prior
7.	Number of Examples	17			

	Aspect of Program	Keep the Same	Increase	Decrease	Specific Suggestion
8.	Level of Information Compared to My Amount of Knowledge	16	2		Push harder

PART II

Please respond to the following statements using the KEY given below.

KEY: SA means you strongly agree with the statement, A means you agree, U means you are uncertain, D means you disagree, and SD means you strongly disagree.

		SA	A	U	D	SD
1.	I found the Tuesday orientation session very helpful.	4	11	1		
2.	The concepts presented in the small group process workshop were helpful.	7	8	1		
3.	I believe I can effectively use the principles of learning and motivation in my own teaching.	10	7			
4.	The clinical teaching technique session was helpful for understanding my own teaching style and preferences.	2	10	4	2	
5.	The session on curriculum development in sports medicine was useful.		2	5	8	3
6.	I have a better understanding of the similarities and differences between allopathic and osteopathic family medicine.	2	7	3	4	2
7.	I can use some of the ideas and skills from the audiovisual workshop.	9	9			

PART III

Please write your responses to the following questions in the space provided.

1. What was the most helpful presentation or discussion during this past week?

_____ on motivation:

One by Dr. _____ on principles of learning;

Dr. _____'s on teaching--both enjoyable and motivating;

Dr. _____'s presentation was the most useful. Had some concrete ideas I can carry with me;

_____'s on teaching;

Principles of learning and motivation;

_____--learning and motivation; _____--group also excellent;

Principles of learning and motivation;

Principles of learning and motivation, Audiovisual materials, Elements of group development;

_____--held my attention, but also had a lot of concrete suggestions;

Audiovisuals;

_____'s presentation on principles of learning and motivation;

All presentations were very good. _____'s was the most outstanding, however;

_____;

Toss up between _____'s and _____'s;

_____;

Audiovisual

2. What presentation or discussion during this past week was <u>not relevant</u> to your needs?

Discussion on DO/MD dichotomy—nothing really new. Enjoyed Dr. _____ about paying for technology—this was on target!

Curriculum development did not deal with the subject specifically on development;

Sports medicine—the presenters spent 1 1/2 hours talking about their program, 1/2 hour trying to talk about how it came about. A 2 hour talk about "setting up curriculums--logistic, organizational, practical aspects" would have been much more relevant to us;

Sports medicine;

Curriculum development;

Curriculum development. The lectures were very interesting and modelled difficulties of developing a curriculum area—but I still don't know if there is a "model" approach to curriculum development;

Curriculum development;

Friday a.m.—D.O./M.D.

Least relevant was probably curriculum development in Family Medicine. True, it was shown that an innovative program was started, but I still have many questions of the how;

None;

All seemed relevant;

The session on curriculum development in sports medicine could have been more relevant. It did not answer many of my questions on the subject;

None—all were good to excellent;

Sports—but don't can them. Rather, help them abstract from their experience more of the general principles of starting a new program;

The sports medicine and the discussion of issues in family medicine. The sports medicine curriculum is exciting, but _____ and _____ seemed to have trouble describing <u>how</u> they developed their curriculum and identifying what principles could be applied to the development of curriculum in another area. I don't think Drs. _____ and _____ have a firm grasp of either the issues facing family medicine (which was to be the topic of this morning's talk) or the allopathic/osteopathic dichotomy (the topic discussed the most);

Sports medicine, allopathic vs. osteopathic discussion;

Curriculum development—sports medicine.

3. What part of the program gave you the most difficulty?

Sports medicine—curriculum planning, logistics of getting the course together was "fuzzy";

None;

Assimilating the large amounts of suggestions/information--I hope I will be able to use most information soon—so far it's all "tucked away";

Parts that are vague, theoretical;

Length of the day in an enclosed room;

Clinical teaching--too much orientation (process) for future activities, not enough content (spread out content over the two weeks);

Coming up with "good" examples during _____'s discussion (practice segments). This was very valuable, however;

Volume/session;

Probably the amount of material and novelty of the material and trying to absorb a good deal of it;

Hard to be so verbal all day—so much new language;

Nothing;

So far, no area of difficulty;

_____;

Nothing has been that difficult so far;

Learning mod(el?) (mode?).

4. What can we do to help you learn during the program?

Provide written handouts for all attending people;

Increase my general knowledge in teaching skills with feedback to see how I am doing. Specifically, I need help on how to answer and ask questions to help students and residents learn;

Provide us with opportunities to practice with feedback--this, I

think, is coming up next week! Perhaps, more small group tasks would be valuable;

More practice, more models or examples to apply;

I think you're doing it well;

Continue workshop and small group functions;

Possibly send out handouts in the summer with relevant leisure reading;

You're already are employing a lot of good educational techniques;

Indexing of readings and outlines;

No specific ideas at present;

Continue the starting and stopping on time with appropriate breaks;

More time trying to apply the ideas--more practice;

Provide exercises and clinical examples for abstract comments;

More active participation.

5. What other suggestions do you have for improving the program?

So far good and no real improvement;

Great as it is. ? more chance for member interaction--sharing of difficulties/successes encountered in our respective programs;

The humanistic orientation is hard to apply. I agree with it and try to apply it. Needs to be more concrete, more examples. I buy into the concept--let's move ahead!;

Possibly a little shorter day;

Expand _____'s teaching time;

Revise and use another approach to curriculum development;

Would like more individual time to discuss projects with faculty (e.g., planning for Friday presentations or major project);

More _____;

So far, I am impressed;

More practical time doing things;

None at this time;

The sessions on group process were difficult to correlate with clinical practice;

More _____.

6. What is your overall reaction to this week's program?

Very good--I have been pleased with the information given so far, but need to apply it;

Very enthusiastic--pleased--looking forward to next week;

Good--glad I came;

Relevant, useful;

Slow start, but I feel well oriented, know all the group by name, and am looking forward to second week;

Outstanding;

Excellent, but overwhelmed by Friday;

Good;

Positive;

Excellent;

Overall, very positive. I feel that it will be very helpful to me in my future teaching;

Excellent--keep up the good work;

Great! This is helping a lot;

Very good--you people do a good job;

Slightly disappointed;

Good except for sports medicine.

ADDITIONAL COMMENTS

Thursday happened to be boring to many people--not necessarily the topics, but the presentation;

Thanks!;

I feel the social events are a wonderful plus to this program. The people running this are sensitive to this;

Would like some feedback on possible ways of handling problems female medical students may encounter;

_____ sessions were very good, but might have been better if there had been less initial discussion, a less-rushed presentation of the key issues, and then more time for discussion of those concepts at the end;

Two particular presentations needed more organization--sports and the DO/MD one. Good idea to have the "social" BBQ the second night--helps establish ourselves as a viable group.

FAMILY MEDICINE FACULTY DEVELOPMENT PROGRAM

End-of-Week Evaluation
Week 2
September 18, 1981

SUMMARY

PART I

Please indicate your overall reactions to this past week's sessions by checking the appropriate box. If you have a specific suggestion about how a change should be made, write that suggestion in the appropriate space or at the end of this instrument.

	Aspect of Program	Keep the Same	Increase	Decrease	Specific Suggestion
1.	Amount of Reading	11		2	Make sure that if a reading is needed for the next day that this is noted; Make more explicit what we are to read; Spread out more; Amount of material OK for further reference; Be more clear on specific readings for each session/-remind us.
2.	Comfort of Room	12	1		E6 is much better--far away from hotel, though; E6 is fine; Change rooms--a.m./p.m. E6 is good.

	Aspect of Program	Keep the Same	Increase	Decrease	Specific Suggestion
3.	Length of Workshop	9	1	3	Gets long at the end of the week; Finish on time; Attend to breaks better; Don't make any longer. 6 hours leads to significant fatigue. Add practicum 1/2 day in 1st week.
4.	Relevance of Information	10	3		Less "student" examples, let's try to focus directly on residents more.
5.	Amount of Participation	10	3		We need more practice trying on different styles; Better! Having residents run "how to ask questions" session was very valuable and should be done more.
6.	Amount of Practice	6	7		We need more practice trying on different styles; Except for practice preceptoring on video, would like more practice;

	Aspect of Program	Keep the Same	Increase	Decrease	Specific Suggestion
					Practice supervisor role more. Better!; Increase time, but try not to decrease information given out; Maybe more relevant practice examples from own experience.
7.	Number of Examples	12		1	Use less videotape vignettes; Fewer vignettes; Change quality, good to increase hospital and residency examples.
8.	Level of Information Compared to My Amount of Knowledge	11	2		

PART II

Please respond to the following statements using the KEY given below.

KEY: SA means you strongly agree with the statement, A means you agree, U means you are uncertain, D means you disagree, and SD means you strongly disagree.

	SA	A	U	D	SD
1. As a result of the psychomotor teaching skills session, I am better prepared to teach these types of skills.	8	5			
2. I will use the approach presented for the teaching of psychomotor skills.	9	3	1		
3. The session on presentation skills will help me improve my own presentations.	10	3			
4. I feel more skilled as a clinical supervisor.	9	4			
5. I will use the ideas presented in the constructive feedback session.	8	4	1		
6. I am more aware of my own thinking as a physician as a result of the discussion on perspectives in learning.		7	3	1	2
7. I found the clinical teaching practice teaching sessions (videotape) helpful.	5	5	1	1	1
8. I have a better idea of how to ask and answer student questions.	5	8			

9. The "practicum" session (Thursday afternoon) should be continued.	10 SA	2 A	1 U	___ D	___ SD
10. I found the practice teaching assignment a valuable learning experience.	9 SA	3 A	___ U	___ D	___ SD

PART III

Please write your responses to the following questions in the space provided.

1. What was the most helpful presentation or discussion during the past week?

Psychomotor skills, presentation skills;

Can't answer—different aspects of many were helpful;

Practice teaching exercises, teaching psychomotor skills, principles of learning and motivation;

Lecture presentation; (_____);

_____;

I liked them all;

Practice teaching on videotape;

Presentation skills;

Those involving our role playing and practice teaching were equally most helpful;

Practice supervisor role. Unfortunately, only done one time. I resent being videotaped and not having a chance to see it. I feel it is unfair and may refuse to participate next time just to make a point. I've no problem with being videotaped. I want to learn from it, not just be a guinea pig for someone's research or tape development;

Teaching psychomotor skills;

Preceptoring role play on videotape;

Teaching psychomotor skills.

2. What presentation or discussion during this past week was not relevant to your needs?

Perspectives in learning was relevant but could have been better if done earlier in the program;

We need a better example of better presented example of curriculum development. A half day more specific on didactics of curriculum development might be more useful;

Curriculum development;

Last week—sports medicine, perspectives in Family Medicine. This week—perspectives in learning;

Role of clinical supervisor already done in my experience;

None;

Teaching history;

Clinical supervision videotaping--some good content, but too much "uninvolved" time and lack of client review of videotape;

All relevant;

Perspectives in learning;

Perhaps the perspectives in learning--could be made more relevant;

Perspectives in learning was interesting and I enjoyed group discussion, but I don't think it changed my behavior. It did increase my awareness on how I think in medicine. Good articles;

I'm still not sure what the purpose of the perspectives on learning session was. It didn't seem particularly meaningful.

APPENDIX K

FIELDTEST DATA: FELLOW INTERVIEWS

FELLOW INTERVIEW QUESTIONNAIRE

NAME COMPOSITE RESULTS

QUESTION #1

Before the September program did you have a background in or any previous experience with:

								Mean*
elements of group development	Y 4	N 10	1	2	3	4	5	1.875
clinical teaching technique	Y 12	N 2	1	2	3	4	5	2.666
role of clinical supervision	Y 10	N 4	1	2	3	4	5	2.3
constructive feedback in clinical education	Y 10	N 4	1	2	3	4	5	2.7
principles of learning and motivation	Y 6	N 8	1	2	3	4	5	2.5
teaching psychomotor skills	Y 9	N 5	1	2	3	4	5	2.44
producing audiovisual materials	Y 6	N 8	1	2	3	4	5	1.83
presentation skills	Y 8	N 6	1	2	3	4	5	2.25
asking and answering student questions	Y 10	N 4	1	2	3	4	5	2.45
perspectives in learning	Y 8	N 6	1	2	3	4	5	1.625

(If yes for any of the above) How would you rate your expertise in this topic prior to the September program on a scale from 1 to 5 with 1 low and 5 high?

*Mean calculated for those who answered "yes" to each of the items.

QUESTION #2

Since the September program ended have you undertaken any additional study in:

								Mean**
elements of group development	Y 2	N 12	1	2	3	4	5*	3.28
clinical teaching technique	Y 6	N 8	1	2	3	4	5	3.82
role of clinical supervision	Y 6	N 8	1	2	3	4	5	3.82
constructive feedback in clinical education	Y 2	N 12	1	2	3	4	5	3.64
principles of learning and motivation	Y 2	N 12	1	2	3	4	5	3.21
teaching psychomotor skills	Y 1	N 13	1	2	3	4	5	3.71
producing audiovisual materials	Y 2	N 12	1	2	3	4	5	3.178
presentation skills	Y 4	N 10	1	2	3	4	5	3.785
asking and answering student questions	Y 1	N 13	1	2	3	4	5	3.39
perspectives in learning	Y 1	N 13	1	2	3	4	5	2.107

*How would you rate your expertise in this topic at this moment?

**N = 14

QUESTION #3

Have you used any of your notes or handouts from the September program since that program ended?

Y 11 N 3

(If yes) For which topics have you used your notes or handout materials?

		HOW OFTEN			
elements of group development	Y 5 N 9	4 once	1 2-3	4-5	5+
clinical teaching technique	Y 9 N 5	2 once	5 2-3	2 4-5	5+
role of clinical supervision	Y 8 N 6	3 once	5 2-3	4-5	5+
constructive feedback in clinical education	Y 7 N 7	3 once	3 2-3	1 4-5	5+
principles of learning and motivation	Y 6 N 8	3 once	2-3	3 4-5	5+
teaching psychomotor skills	Y 6 N 8	once	6 2-3	4-5	5+
producing audiovisual materials	Y 9 N 5	3 once	3 2-3	3 4-5	5+
presentation skills	Y 9 N 5	once	5 2-3	2 4-5	2 5+
asking and answering student questions	Y 5 N 9	3 once	2 2-3	4-5	5+
perspectives in learning	Y 2 N 12	2 once	2-3	4-5	5+

QUESTION #4

Have you shared your new knowledge and skills that you learned during the September program with your colleagues or other people in your organization or community?

Y 13 N 1

(If yes) Which of the following categories best describe how you shared your new knowledge and/or skills? You may choose more than one category if it is appropriate to your situation.

The categories are:

2 (16%) formal presentation

8 (61%) individual consultation

13 (100%) informal conversation(s)

2 (16%) written communication(s)

0 (0%) other (please specify)

QUESTION #5

In the six months since the end of the September program, have you had an opportunity to use any of the knowledge or skills that you learned during those two weeks?

Y 14 N 0

(If yes) Please describe what specific knowledge or skills you have been able to use.

Presentation skills	(11)	Clinical supervision	(6)
AV	(9)	Feedback	(6)
Clinical teaching	(9)	Group development	(5)
Psychomotor	(7)		

QUESTION #6

In the next six months, do you expect to have an opportunity to use any of the knowledge or skills you learned during the September program?

Y 14 N 0

(If yes) Please describe what specific knowledge or skills you expect to be able to use.

Clinical teaching	(8)	Psychomotor	(7)
Lectures	(8)	AV	(7)
Group development	(7)	Clinical supervision	(5)

The next series of questions is concerned with the exercises or simulations that you participated in at MSU that were videotaped for you to review at a later time.

QUESTION #7

Did you review the videotape in which you were placed in the role of a clinical teacher supervising a first-year resident?

Y 12 N 2

(If yes) On a scale from 1 to 5, with 1 low and 5 high, how would you rate your overall perfomance as a clinical teacher in that videotape?

	Mean
N = 12 1 2 3 4 5	3.416

If you were to go through that same clinical teaching simulation tomorrow, how would you rate your expected performance? Again, use a scale from 1 to 5.

	Mean
N = 14 1 2 3 4 5	4.107

QUESTION #8

Did you review the videotape of your presentation assignment? If you remember, that was the one on the last Friday of the September seesion where you were asked to teach something to someone.

Y 12 N 2

(If yes) On a scale from 1 to 5, how would you rate your overall performance in that presentation assignment?

		Mean
N = 11	1 2 3 4 5	3.227

If you were given the same assignment, to teach something to someone for twenty minutes, and you had to do it tomorrow, how would you rate your expected performance?

		Mean
N = 13	1 2 3 4 5	4.192

One respondent found this "hard to rate" and did not provide a response.

QUESTION #9

Did you review the videotape of the research and evaluation project presentation that you gave in January?

Y 10 N 3

(If yes) On a scale from 1 to 5, how would you rate your overall performance in that presentation? Note that the focus is on your presentation and the associated skills, not on the content of the research or evaluation project that you presented.

		Mean
N = 10	1 2 3 4 5	3.45

If you have to give a similar presentation tomorrow, how would you rate your expected performance?

		Mean
N = 12	1 2 3 4 5	3.875

One fellow did not complete this assignment.
One fellow did not give a point rating for the second question.

QUESTION #10

Has your participation in the September program changed your role or function in your organization? For example, have you tried some new teaching techniques or have you significantly changed any of your daily activities?

Y 7 N 7

(If yes) Please describe how your role or function has changed.

Trying to do more or new types of teaching (4)

Not doing much teaching right now (2)

QUESTION #11

Has your perception of teaching as a career changed since the completion of the September program?

Y 9 N 5

(If yes) Please describe how your perception of teaching as a career has changed.

More interested in it (3)

More comfortable, confident (3)

View teaching as more of a science, less than art (2)

QUESTION #12

Do you feel that the September program has helped you to become a better teacher?

Y 14 N 0

(If yes) Please describe how the program has helped you become a better teacher.

(If no) Please describe why the progam has not helped you become a better teacher.

Gave me conceptual framework to use when teaching (7)

Gave me information and skills I lacked before (5)

Gave me practice on my presentation skills (3)

QUESTION #13

If a friend or acquaintance of yours was interested in becoming a faculty member in family medicine or wanted to become a better teacher of family medicine, would you recommend the September program to him/her?

Y 14 N 0

(If yes) Why would you recommend the September program to someone interested in becoming a faculty member in family medicine?

(If no) Why wouldn't you recommend the September program to someone interested in becoming a faculty member in family medicine?

Can learn things that are useful as teacher (5)

Physicians have little exposure to educational methods (4)

Gives you a structural framework for teaching (4)

Recommended it already (3)

QUESTION #14

Is there anything else that has happened to you as a result of the September program that has not been covered by these questions?

Y 4 N 10

(If yes) Please explain or describe.

No common response given.

QUESTION #15

Do you have any additional comments or concerns that you wish to express at this time?

Y 8 N 6

(If yes) Please make them at this time.

Don't understand where perspectives in learning fit in (3)

Concrete, practical things most helpful (2)

APPENDIX L

FIELDTEST DATA: FINAL DEBRIEFING

FAMILY MEDICINE FACULTY DEVELOPMENT PROGRAM
1981-82 FINAL DEBRIEFING

SUMMARY

PART I: Written Responses

1. What is your overall evaluation of the program?

- Overall program was worthwhile--some half-day sessions weren't worth time and effort travel, etc. Highly recommend program to those in or going into Family Practice Residency programs as instructors.

- Very useful--probably will prove of even more usefulness as I get more involved in the teaching of family medicine. I'd recommend it to a friend who is serious about wanting to teach family medicine.

- I feel that the program was valuable for me as a future family medicine educator and that I have learned a great deal of relevant information that will be useful to me in the future. I would recommend it (and have).

- Yes (would recommend/worth my time). Hopefully next year we can continue to have individuals from our school attend this fellowship.

- Worthwhile for those interested in teaching medicine at any level.

- I would recommend it to a friend who was definitely committed to a future in academic medicine. Those with only interest in part-time teaching did not need so extensive a course. It was worth my time if I did not have to travel so much.

- Very worthwhile. Generally covered areas not covered at all in routine medical education. Good use of people from other disciplines than medicine who's expertise is very relevant to our tasks but in general who were familiar enough with the peculiarities of the medical system to be relevant.

- I enjoyed the program a lot and would recommend it to others.

- Overall it was a good program with real applications for any further teaching responsibilities I might have. The first and last sessions were of most benefit.

- Excellent--met my goals. Would (and have) recommended it to others.

- Sept: Very good. _____'s programs were helpful, some of clinical teaching with _____ were good. _____ and

_____ give good programs on sports medicine, but bad programs on curriculum development. Jan: Research presentations were +/-, could have been better. March: _____ was good, excellent, _____ terrible. May: _____ excellent in May; mediocre (due to dull topic) in March. We didn't care about the structure of TV cameras of his program.

- Overall, the program was excellent. I would recommend it to any family practice resident, regardless of whether or not the person was pursuing a faculty position in family practice. Teaching skills are applicable to many environments. Exposure to grants and research will increase substantially, the likelihood of myself doing something like that.

- Excellent. Yes (worth my time). Have already recommended it.

- Overall evaluation and reaction to the program was favorable. I feel that it was well organized, well presented, and that the topics were pertinent. I would recommend the program to a friend who might be interested in a career in family medicine. It was worth the time, though I do feel the same amount of material could be presented in a shorter period of time.

2. What was missing most from the program?

- Specifics--a lot of conceptualization was done without getting into actual solutions--more concrete answers; granted that there are some.

- I felt the marked contrast of being here and 100% involved in teaching education, then being back home and 100% involved in being a resident was a disappointment. I know you've identified this problem, arranged site visits, made projects relevant to home institution. Still, the contrast persisted and I don't know if anything more could be done about it. Maybe if I hadn't been still a resident...but I'm grateful for the opportunity.

- Perhaps the only thing that I feel was missing was more clinical relevance to the model of medical education used in my program (which is more attending than active preceptoring); although I'm not sure this affected my learning.

- I was the only person involved in a situation not related to Residency Practice planning and mostly felt left out.

- Much of my time is spent on teaching rounds at the hospital. This was totally ignored in the program and really needs to be addressed. The one lecture on presentation skills was useful, but much more time needs to be devoted to this topic. I gave several lectures this year and needed more work on presentation skills. Therefore, two areas that are predominant in teaching were not

covered adequately. You should also include information on clinical teaching with small groups, i.e., conveying information during rounds while maintaining efficiency in patient care. You need to get into the specifics; that's what's really difficult. The theoretical perspectives by _____ were useless to me.

- Hard to find time and energy to apply many of the things we learned back in the work setting.

- How to teach residents when backup at the outpatient clinic. Computers. Make-up and structure of a residency program. Books or good reading sources (not a voluminous list--make it short).

- Some of the more practical aspects of being a faculty member, e.g., preparing budgets, interviewing candidates for residency slots or faculty positions.

- Good examples of research and grants. Too much time picking apart those with problems rather than seeing one good example as a model. Would also like more time for interchange of ideas and discussion about individual programs, problems, and innovations.

- Would have liked a little more on political issues. Understanding State and Federal financing into Family Practice.

- Developing curriculum in family practice was weak, as was program evaluation and funding of family practice. Most speakers were not family physicians--but this did not reduce their credibility, as they all seemed to be very tuned in to family practice.

- Perhaps too idealistic, not enough of how to deal with people who haven't been here; not enough about dealing with the hopelessly incompetent student (or adminstrator).

- I think expectations were met with the sessions on teaching. An area where expectations were not truly met was in the area of medical writing.

3. What comments do you have about the administration of the program?

- Tough to get out of residency program for 5 weeks without some resentment--assistance at MSU better and more available than that at home office. Tough to get to East Lansing twice in the winter. Maybe schedule March or January session in April. Logistics of leaving Friday afternoon to make flights, etc.

- Well done. Schedules mailed out prior to sessions and checklists were very useful. Spreading out sessions thoroughout the year was good and worth the hassle involved working around rotations. My program did not obstruct my coming here--I felt it was very important for my education and career.

- No real comments--overall, a well-administered program.

- Schedule was flexible and we appreciated it. Assistance was adequate and always there.

- Could it be condensed into less time? On the other hand, the practice sessions were most valuable--we needed the chance to practice what we were learning.

- Well administered. Have one-hour lunch breaks. "Night out" not the night before presentations. On-site visits and assistance were adequate and I have no complaints.

- Because of living in Lansing, I think I assumed, as did faculty, that there would be plenty of time to meet and discuss the major project, etc. But in fact we only managed to meet once. I think it's probably important for local fellows to still be given meeting times during the fellowship week (e.g., with advisors about projects). I suspect site visits have the positive effect of rekindling enthusiasm for the project.

- The visits were excellent and I received all the assistance I desired. Schedules and keeping to it were well done.

- The first two weeks took a chunk of time from that month's rotation. I'd suggest breaking it up. I'd also suggest scheduling the sessions for either the first or last week of a month. Sessions that occurred in the middle of a month disrupted that month's rotation.

- No administrative problems that I was aware of.

- Fine!

- Program was well administered while physically at MSU and also while away, by mail. Assistance was available when necessary, around the major project and also around scheduling time to make it to MSU. Making a "book" of the major projects was good.

- Shorter lunch (1 1/2 hours optimal). Overall well run. I liked having the teaching sessions first so that I could evaluate subsequent presentations based upon those principles.

- I feel the adminstrators made a conscientious effort to run a very efficient program. It was nice to have assistance available to the fellows throughout the fellowship. I had a little difficulty with the fellowship schedule because of commitments to my primary job.

4. How would you rate your contribution to the program?

- I hope I contributed my share. I feel I was vocal in all the study groups, etc. One can always be more involved--I learned a lot during breaks and after hours in informal discussions with fellows. I feel I should have taken advantage of OMERAD facilities more.

- Yes (I was involved as I wanted). Yes (I took advantage of fellowship opportunities).

- I believe that I was involved and took advantage of the educational opportunities, participating actively in both the fellowship sessions and the after-hours sessions where a good deal of informal learning was available.

- Participation would increase with more diversity of topics. Involvement could have been more if home-base situations were more understanding.

- I was average in my participation. I thought that participation of the fellows was fine.

- Ideally I would have liked to do a project more directly related to my residency, but my needs were fairly clear to complete and write up my research project. I think I could have applied more of what I learned on a different kind of project. (However, I was very grateful to have help doing what I did!)

- I rate my contribution as average to above average and received feedback from this.

- I probably talked too much.

- More "take-home" assignments. Not so much for my experience, because I think I did--tried a lot back home, but maybe others would benefit from being pushed more to try some of the stuff.

- I felt involved and as though I had something to add. Specifically, "allowing" the fellows to experiment and be comfortable in the group and discussing group dynamics was helpful.

- I occasionally felt I wanted to participate more but wasn't sure of how to do so.

- I would rate my contribution to the program as average. I feel that because of outside responsibilities, I was probably not quite as active as I would like. Also, since I had just obtained a new faculty position, I do not feel that I was comfortable enough with my new duties to be able to fully tap the fellowship.

5. What would better prepare fellows for the program?

- I feel any fellow in a residency program, preferably 3rd year or any attending in a F.P. residency program, is "prepared" to take the program.

- Probably most important is a fellow's commitment to teaching. Having a group full of committed educators will assure active and fruitful participation.

- Reading list or expected topics before the program starts--could help us cover more material.

- Better orientation materials before coming to session, describing the details of the year.

- Talking to former fellows.

- Possible reading sources before the program or articles to read.

- Discussions with previous fellows (I did this and I think I had a pretty good perspective on the program before I applied).

- Everyone should be well versed in its content--expectations prior to coming to get the most out of it.

- A 2-3 page handout describing some of the main areas covered in the program, i.e., teaching skills, research, grants, implementations, etc. Mention small group work, major project, understanding AV, etc.

- I think I would better prepare my program director for what the fellowship had to offer, to help the fellow use his/her residency as a more open "lab."

- I think future fellows might appreciate a preview of the sessions that will be offered over the course of the program before they start it in September.

6. What comments do you have about the evaluation of the program? (pre/posttest, telephone interviews, end-of-week evaluations)

- All are necessary, and I hated doing them. I'm the type who would like to know in advance what I am responsible to do.

- Pre/posttests frustrating. Sometimes they violated _____'s principle--"tell them what you want them to know." Knowing "the elements of group development" hasn't proved relevant to my job as a teacher. After 9 months, I don't know how I did on pre/posttest

(which took time) or what it means. Telephone interviews made me feel bad that I haven't used many of the handouts. End-of-week evaluations were useful. You obviously use them and are open to suggestions.

- The pre/posttests were, in my opinion, worthless. The telephone interviews probably gave a better idea about what was being used, although I'm not sure about the outcome. The end-of-week evaluations were probably the best of the three. A long-term (one-year? five-year?) followup should be done.

- Pre/posttests--too many, but necessary. Telephone interviews were alright. End-of-week evaluations were understandable and necessary.

- The pre/posttest demonstrated that I did not memorize what the presenters felt important. In fact, without studying the material first, the test was difficult to answer. End-of-week evaluations and telephone interviews were fine.

- Pre/posttest--tedious, felt like a guinea pig for someone's study. Also some test times required knowledge of jargon not familiar to me. Telephone interviews were okay--more useful than pre/post-test. End-of-week evaluations--a necessary part; handing them out Monday seems like a good idea, especially when there are numerous different sessions and presenters.

- Do not give the pre-test when we first get to program--it was a turn off for me. Wait a day. Hopefully will get feedback about the final evaluations of the program. Enjoyed getting the feedback results we have gotten.

- The pre/posttest is jargonistic and if you plan to continue to use it, it should be rewritten. I think that the spontaneous suggestions following the individual sessions are a more reliable source of evaluation.

- Pre/posttest was a big pain; hope it was necessary. End-of-week evaluations were most valuable.

- Evaluate briefly on closer to a daily basis.

- Evaluation was extensive, not too disruptive--as time was allowed for the evaluations to be completed.

- Suggest followup 1-2 years from now with a phone interview similar to the one already done.

- I feel that it was an improvement to hand out the end-of-week evaluation at the beginning of the week. Pre/posttests were a lot of work. Telephone interviews--did not mind them.

7. Comments

- Of most benefit was fellow fellows daily discussions in response to educators' lectures and small group experiences.

- Nice setting--both classroom, campus, and outdoor recreational activities. I'm sad it's over--I'll hope to keep in touch with many of the fellows.

- More out-of-state fellows who are definitely invested in an academic career. The stipend did not cover or barely covered my expenses; as compared to local people who had little or no expenses. That should be taken into account when determining the amount of the stipends in the future.

- Should, for practical reasons, have it in three sessions: 2 weeks in September, 2 weeks in February, and 1 week in May.

- I would like to know ways I can further my faculty development after this course.

- A few sessions could be eliminated or strengthened (e.g., curriculum development could be more varied, practical; Dr. _____'s could be eliminated). Would a portion on use/abuse of PA's, nurse clinicians, other health professionals in practice and teaching be valuable? Add session on audits--this is an important committee function--basis for research and looking at problem areas. How to successfully take charge or partake in an audit.

- The D.O. and M.D. mix in the program added interesting perspective to the educational process of the fellowship. A non-bias discussion group (too much political intrique with the session in the first week) around where the two professions fit into the health care system would be of interest--especially some issues around manipulation.

- More practical aspects of how to help impaired physicians. Less on documenting their existence. Also, would have liked the OMERAD faculty to make more comments during Major Project presentations (although I appreciated your allowing the group to make comments). Kind of disappointed there.

PART II: Discussion Topics

1. Major Projects

- Site visits helpful
- A positive, worthwhile experience
- Progress reports, presentations, and notebooks helpful
- Learned from hearing other fellows' presentations
- Politically beneficial to sponsoring institution

Suggestions:

- Incorporate minor projects into major projects
- Smaller groups to discuss projects
- More feedback
- Don't require contract/signature from residency director
- Make clear to director what is required of fellow and what they will receive in turn

2. September Session: "Teaching and Learning"

- Mixed reaction on length of session
- Avoid pretest on first day
- Structure of first session good--group interaction beneficial

A. Elements of Group Development:

- Highly theoretical--hard to grasp
- More practical clinical examples

B. Clinical Teaching Technique:

- Did not use profile instrument
- Too medical student oriented; not designed for this audience
- More practical application to hospital teaching rounds
- Focus on resident teaching model, not preceptor model
- Clinical teaching simulation helpful

C. Curriculum Development in Family Medicine

- Drop _____ and _____
- Give basics of curriculum development and have fellows discuss their institutions specifically

D. Issues in Family Medicine:

- Drop
- Fertile topic, useless session

E. Producing Audiovisual Materials:

- Need more time
- More medical photography

F. Presentations Skills:

- More in this area, including public speaking, professional meetings, lecture/interaction presentations
- Add small group teaching skills

G. Perspectives in Learning

- Drop
- Too theoretical; focus on practical concepts/skills

3. January Session: "Research and Evaluation"

- Need good example of research and good questionnaire

A. Direct Observation and Rating:

- Topic useful
- Presentation weak
- TV not appropriate

B. Issues and Strategies for Clinical Evaluation:

- Generally negative feedback
- Needs to be more generalizable

C. Planning and Conducting Research in Applied Settings:

- Good information but poor presentation
- Include people who have done research

D. Program Evaluation:

- Least helpful
- Need more activity
- Show good as well as bad examples

E. Questionnaire Design:

- Expand into practice (lecture/AV)

F. Writing for Publication:

- Good content
- Disliked exercise
- Use materials that were requested to be brought

4. March Session: "Issues in Family Medicine"

A. Time Management:

- Useful

B. Funding of Family Medicine:

- Generally positive feedback
- Be sure to get national focus

C. Committee Membership:

- Generally positive feedback
- Get perspective of hospital administrator

D. Grant Writing:

- Session well done
- Activity frustrating--many points not clear
- Need good examples to clarify

E. Health Policy and Planning:

- Generally positive feedback

5. May Session: "Administrative Skills"

A. Administrative Skills:

- Discuss dealing with different administrators
- Include budgeting basics and skills
- Perhaps have an additional speaker

B. Hidden Curriculum:

- Relate to family medicine educators
- Interesting information
- Think about a session on stress management

APPENDIX M

FIELDTEST DATA: PROGRAM DIRECTOR INTERVIEW

Interview Responses of FMFD Program Directors
6/16/82

KEY: A - Program Director A
B - Program Director B

QUESTIONS #1 AND #2

B: The number was easily very close to being twice as large as ever before and this had implications for physical things, the room, how are we going to monitor these people.

A: We expanded the clinical teaching component and one of my concerns was whether or not those three sessions held together as a unit. How much was repetitive, how much was new, how much was consistent? More time, new content, added clinical teaching technique content. This was expanded in response to previous evaluations.

B: I knew that perspectives in learning was a high risk going in.

A: The mix of M.D. - D.O.'s a concern. M.D.'s mostly third year residents; D.O.'s both residents and faculty, mostly faculty. Thus they found themselves in different teaching situations. Both would be doing clinical teaching, but the D.O.'s more likely to be doing formal classroom teaching.

B: Mix-up between major project and assignments in the past. This time we stipulated the separation between the two from the very beginning. One thing I felt was useful at the end.

QUESTION #3

B: My major surprise was the fact that the general class of activities was less well received than ever before and the fact that we lost a fellow due to _____. It did not go over well.

A: _____ 's clinical teaching stuff perceived better in earlier offerings. Quite surprised, they said it focused more on med students, preceptor model, rather than on teaching residents.

Another surprise was that despite all our advance work and proscribing content and behavior, _____ and _____ talked about sports medicine rather than curriculum development.

B: Along with _____ and _____, they pointed up the weakness in using clinical personnel as faculty in this program.

A: We need to use them, but they're not pulling the freight. Also surprised by the reception of the AV workshop. They wanted more of it, wanted to actually produce the stuff.

B: We both came out with the notion that this is a nuts and bolts group. The group we had last year was more academic, liked topics of a general nature. Yet, I wasn't less satisfied than last year. We had a different group and we knew right then that certain things would go over well for them and that it required a different curriculum. Don't really know what happened with _____ , either.

A: _____ was coming off a year's sabbatical and maybe wasn't as well tuned in as before.

Disappointed in the skill and motivation level of the _____. Ones we had before I thought were of higher quality. _____, _____, _____ were very good. This group couldn't apply information, just rote recitation. Disappointed in their ability to verbalize what we were trying to teach.

B: I agree with that, also with knowing that that was the group that was encouraged by directors to come as opposed to they initiated the request. Looking back that will make a tremendous difference in how we recruit classes in the future. You can't force somebody to like teaching.

QUESTION #4

B: Difficult to say who stars will be. Easier to identify the others, the losers. I knew, I could tell by the type of questions they asked, how much they knew about the program. When I visited _____there was already some hesitancy on his part. _____ and _____ were both obvious risks, we knew _____ already and _____ who was using the fellowship to pump up his own program and to get information to use for his own program and grant. _____ was ambiguous about what the fellows could do for him while _____ was very specific, positive about possible projects.

A: I had some similar experiences and had some surprises. Not impressed with _____. Didn't expect as much as we got. Same with _____. Met him in social setting, that colored my reactions. Same for _____, obvious he wouldn't be superstar. Question just how far he could come. _____ also obvious from day one that he would be a problem.

On the other hand, guys like _____, I knew he was good as soon as I walked in the door.

B: Same with _____.

A: With one or two exceptions it was pretty easy to do. One thing that was disappointing, we needed more of a clarification of how many of these people were going into full-time teaching. Had several go into private practice which was different from their applications.

B: We are pushing the next group harder to be honest with us as to how they're going to use it.

QUESTION #5

A: No baseline to judge it by.

B: General impression was that they had a long way to go. They had been exposed to it, but were not very polished.

QUESTION #6

A: Easy for me to see changes from September to January for _____, _____, _____, _____.

B: _____ struck me as one really beginning to put things together.

A: _____ regressed, the focus was on research, he wasn't really sure what he was talking about, hadn't really done much in the way of presentations before. _____ improved immensely.

B: He started to spark in my mind at that time.

A: He conceptualized a good evaluation scheme and laid it out in a well organized manner. I was somewhat disappointed with _____.

B: He's hot and cold.

A: _____ had no apparent carry-over.

B: _____ has no preparation. Except for the final major project presentation. He would not bring up anything that indicated preparation.

A: Some individuals used presentation principles we talked about while others that watched them didn't pick that up or chose not to do likewise.

B: I saw a lot of them make attempts to use overheads and some organization that I had not seen before. Some modeling of that. Some pretty primitive, violated rules, but got sense of how to use AV. Became slightly more polished.

A: Some improved, as a group saw about 50/50 improvement.

B: More like 70/30 improvement. 70% showed improvement. I saw the really bad ones in September. _____, _____, _____.

QUESTION #7

B: The major project presentations were the most well thought out statements they had made. Mainly because we pushed them. Reinforced in their minds what they would be presenting, that they would have to have handout materials. Motivated them to sit and think about what they were going to say. Some had already given their presentations to other groups. Had some experience already presenting.

A: As a group they were much better than January.

B: I'd agree with that. They felt that they had made improvement. They were proud of their efforts. There were few apologies made about the presentations or their papers. Whereas I had a lot of apologies in the earlier sessions.

_____ was a negative surprise for me. When I first met him he was always enthusiastic and asked questions, but there are still some basic teaching skills he's lacking.

A: _____ came in a little lower than my expectations. I had higher expectations for him than I saw.

B: I had two expectations for _____. _____ the person as he participated in the program and the other _____ and what he accomplished with his major project. He came of his own volition and made a good permanent contribution to his program. But he didn't get involved as much as I would have liked.

QUESTION #8

B: _____demonstrated understanding of how much information to put on slides.

A: No specific examples, but I do remember people either in jest, in lighthearted manner talking about how someone else violated principles or didn't do this or that. _____ for example didn't use overheads or handouts and they commented on this.

B: A couple of times with _____ we went over basic principles of overhead design. _____ called about developing guidelines for supervision, referred to things _____ talked about.

QUESTION #9

A: No.

B: I gave out information on perspectives in learning to and _____, gave them an extra article I had not given to the entire group. With the presentation skills they wanted more practice, not more content.

QUESTION #10

B: _____, when I went out to work with him on the major project as an aside he asked me to help him with two major presentations he was giving. We sat down and discussed presentation skills, audience involvement techniques. He called us and as a result of that he picked up three more lectures to give to first year students. And he brought in his evaluations and they were excellent.

A: I don't know if we can classify submission of STFM papers as unintended outcomes, but was pleasantly surprised that several people had done that.

QUESTION #11

B: When you look at all of them, they're still a little bit better than past groups. A couple are clearly low, but there's a couple of those in each group. I think they came in lower. I think we did more for them to clean up their acts. I think they went farther and ended up higher.

A: I disagree. I would classify, group wise, the overall entry skills of this group as higher than last year. I think we had a higher quality fellow this year, didn't take them as far, but overall quality this year is higher than last year.

B: I think the presentations and major projects were clearly better than last year or any other year.

A: I agree.

B: Would have to do analysis of each one. Numbers deceive you too. More bad apples this year.

In terms of taking advantage of what the program had to offer, they played around with their projects, nothing will change in their lives as result of being in our fellowship. _____, _____, _____. _____, I think picked up some skills, but can just write those other three off.

A: _____ came in at higher level, he was a former teacher. _____ right up there, _____, _____, _____.

B: _____ also former teacher. _____ did well.

A: _____'s presentation not great, neither was _____.
_____ tried hard but had not quite put it all together yet.

B: Hard to separate presentations from quality of the project.

QUESTION #12

B: _____ is a counter example of what I originally thought. Overall I was happy with the group, didn't have major concerns, little disappointed in their inability to generalize or translate any information we gave them. If we didn't have a practice session and talk about specifics they gave up pretty quickly.

A: We had to make the application for them. They couldn't.

B: There were times the presenters were not what they wanted and they got passive. I was a little disappointed in that.

A: We have yet to underestimate their abilities. By successive approximations we're coming closer to a better program. I'm not going to throw out anything based on the reactions of this group. Will make curricular changes, do that all the time. It's more that we learn as they do.

B: We have to be careful taking their opinions of what they say they can do.

APPENDIX N

FIELDTEST DATA: SUPERVISOR INTERVIEWS

SUPERVISOR INTERVIEW QUESTIONNAIRE

NAME_____ COMPOSITE RESULTS

QUESTION #1

Once Dr. _____ learned about the FMFD Program, did you encourage him/her to participate in the program?

Y 14 N 0

(If yes) Why?

(If no) Why not?

Dr. _____ had expressed interest in teaching as career	(6)
I felt Dr. _____ needed some skills improved	(3)
I saw Dr. _____ as a potential faculty member	(3)

QUESTION #2

Has Dr. _____ shared any of the information or new knowledge or skills that he/she learned about teaching during the September program with you or other members of your organization?

Y 12 N 2

(If yes) Which of the following method or methods best describe how he/she shared this information?

3 (25%)	formal presentation
4 (33%)	individual consultation
12 (100%)	informal conversations
1 (8%)	written communication
0 (0%)	other (please specify)

QUESTION #3

Do you know if Dr. _____ has been able to use any of the new knowledge or skills related to teaching that he/she learned in September at MSU?

Y 15* N 0

(If yes) Please describe the types of knowledge and/or skills that Dr. has been able to use, and the types of situations that they have been used in.

Clinical teaching skills (5)

Presentation skills (5)

Group discussion skills (4)

QUESTION #4

Have you observed Dr. _____ doing any teaching since late September? This could include activities such as one-on-one clinical teaching or supervision, small group discussion teaching, or formal lectures or presentations.

Y 12* N 3

(If yes) How often have you observed Dr. _____ doing some teaching since late September?

1 once

1 2 to 3 times

2 4 to 5 times

8 more than 5 times

Clinical teaching (5)

Presentations (5)

Group discussions (3)

QUESTION #5

Do you feel able to judge whether or not Dr. _____'s teaching behavior has changed since late September?

Y 12* N 3

(If yes) How has Dr. _____ 's teaching behavior changed since September?

More confident, comfortable (5)

More organized (3)

What do you think has caused the change in Dr. _____'s teaching behavior?

Exposure to program (2)

More mature (2)

*Two different supervisors were asked questions 3-5 for one of the fellows.

QUESTION #6

Have you noticed any change in Dr. _____'s role or function in your organization since the end of the September program? For example, has he/she become active in new areas of your program or has he/she taken on new responsibilities?

Y 10 N 4

(If yes) Please describe this change in Dr. _____'s role or function.

Trying to do some research (2)

Serving as coordinator of preceptor program (2)

QUESTION #7

Has your program benefited in any way by Dr. _____'s participation in the FMFD Program?

Y 13 N 1

(If yes) Please describe how your program has benefited.

(If no) Please explain why you do not believe that your program has benefited.

Has shared new information with us (3)

Has expanded our educational base - content wise (3)

The major project (3)

QUESTION #8

Would you encourage another resident (or faculty member) from your program to participate in the fellowship program in the future?

Y 11* N 0

(If yes) Why?

(If no) Why not?

Opportunity for resident to check out academics as a career (2)

Opportunity for young faculty to grow (2)

Trying to accelerate development of a young program (2)

*Question not repeated for three supervisors with more than one fellow in program.

QUESTION #9

Do you have any additional comments about either Dr. _____'s teaching behavior or skills or about the FMFD Program that you would like to make at this time?

Y 11 N 3

Logistics of fellowship very workable (2)

Would be helpful if supervisors know more what they can do to help (2)

QUESTION #10

Would you like to receive a copy of the final evaluation report on the FMFD Program?

Y 11 N 0

QUESTION #11

Do you have any other comments or concerns that you would wish to express at this time?

Y 2 N 9

Comments combined with #9.

APPENDIX O

METAEVALUATION DATA: PROGRAM DIRECTOR INTERVIEW

METAEVALUATION DATA: PROGRAM DIRECTOR INTERVIEW

The evaluator conducted a telephone inteview with the two directors of the FMFDP on Friday, October 22, 1982. The information gathered during this interview was used to answer four research questions as part of the metaevaluation of the fieldtest of the evaluation framework. A transcription of that interview follows.

KEY:
A - Program Director A
B - Program Director B
E - Evaluator

RESEARCH QUESTION #2: Was the evaluation framework practical in its use of resources?

SPECIFIC QUESTION: Did the evaluation procedures produce information of sufficient value to justify the resources expended?

B: I felt, in general, yes. But I think that there were specific sources of data that were expensive and time-consuming for everybody and we got very little out of, in particular the pretest and posttest. It could be a weakness in the instrument, but I felt that there wasn't a lot of information that came out of it and a lot of time went into getting the faculty to write the items, going over rewriting the items, scoring the items, and it took a fair amount of time for the fellows to complete it. I just kind of discounted the data at the end. That for me was one of the major weaknesses.

A: I guess I wouldn't come down as hard on the cognitive assessments, but they were a pain in the chops. But I think that like any test there's always a high initial cost. What I've been thinking about, we did it once, would it be worth doing every year? And my answer to that would be, "No, I don't think so." I felt justified in using the resources that we did for a one-shot deal. I don't think that it would be worth repeating all of those every year. But I think they did provide some valuable data that we just didn't have.

B: One of the reasons I felt the paper-pencil test was weak was that the main emphasis of those two weeks is heavy on skills that are assessed in some form of demonstration. I think it's real hard for faculty to write items that assess whether someone has identified a skill or not. I felt that is where we kind of fell down. That's why I say that it might not be the method that was bad, but the fact that if we are going to continue doing this that we need to come up with items that are better written and more valid given the skills.

SPECIFIC QUESTION: Were the evaluation procedures administered so that program disruption was kept to a minimum?

A: I didn't see that it intruded much at all, the only possible exception is the cognitive pre-, post-, delayed posttest, but other than that we're doing most of those things anyway during the actual two weeks of the program. You did a lot of things like your telephone calls and that outside the program didn't really disrupt the program, so no, I don't think the precedures overall were that invasive at all.

B: I pretty much concur on that. Are you interested in knowing if it had been disruptive if we had been doing the evaluation? Or only if it was conducted with an external person doing it? If I personnaly had been conducting it I think it would have been very disruptive as to running the other parts of the program, trying to do both of these simultaneously.

E: Do you think the results would have been different as well?

B: Yes, I think they probably would have been.

A: How?

B: Well, it's hard to speculate, but usually when someone who has designed and delivered part of the program they become so invested in what it is they are evaluating that it becomes hard to be totally unbiased. My guess is when I spoke with directors and conducted some of those other activities that the results might have been different.

A: I'm not so sure. I think that anybody who's doing an evaluation or who has an educational background would probably have gotten the same story.

SPECIFIC QUESTION: Did the use of multiple instruments appear to yield results that justified the extra time and effort involved in their development and administration?

B: My inclination is that because it was a more comprehensive study that it was definitely justified for doing it the initial time. I think one of the things you learn is where do you get redundant information and then you can ask yourself, "Given that you want a certain quality of information, which source might I pick?" If everything is just confirming what you hear over and over again you can start to select out that procedure that gives you the best information for some of the least effort. I think that all the

telephone calls plus the debriefing at the end as well as the End-of-Week Evaluations, we were starting to get some redundancy in our information. If I were to do it again, I would probably chop out parts of that.

A: I basically agree.

RESEARCH QUESTION #3: Was the evaluation framework valuable in providing information to you as decision makers?

SPECIFIC QUESTION: Did it provide information that answered specific questions that you had about the program?

A: I think it did. At the time we were in our fourth year in operation and there were some nagging questions that I always thought about, but because of the complexity had never mounted a sufficient evaluation effort to answer those questions. In a sense your framework answered those questions more in a confirmatory note. I had sort of hunched, but now I had data that told me that we were doing the right things or we were way off-base here, that was very helpful to me. Again, the question comes, would I do this every year, and my answer would be "Probably not." Maybe if there were changes in the structure of the program, if we went to another format or if we had significant changes in the faculty that presented during the session then I would want to repeat. But as the program stabilizes, now that I have this information I feel much more secure in knowing that we have some data that says we're doing something right.

B: I'd agree with that.

SPECIFIC QUESTION: Was the information that you received complete and comprehensive? Was there anything left out that you would like to have known?

A: Perhaps the only bit of information that was missing, and I don't know how in the world you ever could have collected it, was some documentation of them using the skills in their setting. In other words, they told us they do lectures and work with medical students, but if we could have actually observed them performing some of the skills taught during the September session in their own environment, that would have been very helpful to me.

B: I would agree that it's always ideal to see if there is direct application data. There were times when if anything I felt that I was maybe inundated with information. In a lot of individual

information that was useful, but I might have wanted to see something a little more summarized. But that's more preference that anything else.

A: There was no problem with what I got, but as an administrator and having a certain level of comfort with the program as it si, I would like to have been able to get a one- or two-page executive summary of just the bottom line.

SPECIFIC QUESTION: How could the evaluation have been changed to provide more useful information?

A: Maybe changing the cognitive level of the test items, making them more application and problem-solving rather than recall.

B: Yes, we really loaded them with that, I think that was a function of writing them kind of at the last minute.

One idea, just for organization, given that it's a program evaluation and really a fellowship evaluation where you look at the whole thing, maybe it would be a useful summary to look at the major goals of that two-week session and the reason I'm saying goals and not objectives is because you get so bogged down once you get into all 35 objectives, but maybe if you looked at the major goals and determined if we have evidence that they were attained or not or to what extent, may be another way of organizing it. I'm not so certain that would have been better, but it's something to think about since that was really the macropurpose from our point of view, "Were our goals reached?"

A: Another possibility would be to get some baseline data on the performance skills, like how well could these people do presentations before they got there and how well could they do clinical teaching before they got here. I recognize some of the problems in doing that, but in an ideal it would have been nice to have some entry level behavior.

B: Yes, I've wondered about that, too. Another thing that would be ideal, but would also put contraints on the fellowship, is if we could somehow standardize the presentations so that they were more equivalent to begin with. In other words, we give them a lot of freedom to do different sort of activities and if we were to give them more guidelines,...

A: Proscribe it a little more...

B: ...It would be easier for us to determine if they have really been applying some of the things that we have been talking about, but at the same time they would lose the freedom of picking what they want to do.

RESEARCH QUESTION #4: Were the methods/instruments used within the evaluation framework technically adequate?

SPECIFIC QUESTION: Were the sources of information described in enough detail for you to assess the validity of the information they provided?

A: Yes, I'd agree because I think we're familiar with it. The question would be more appropriate for someone who isn't as familiar with the program as we are. I've got no problem with it, but I don't know how someone who isn't intimately familiar with the program design would respond.

The point is that we worked with you to help develop some of these things so we knew from the very beginning where you were going, but the question as I understand that you are asking it is if in the report, in the summaries, is there sufficient information to assess the validity of the instruments. My answer would have to be "yes" because we were involved very closely, but I don't know if somebody else reading this report like my project officer in Washington would be able to. I would say for me that it's no problem, for others I just don't know.

B: I would agree.

SPECIFIC QUESTION: Were the information-gathering instruments and procedures described in enough detail for you to assess the reliablity of the results they produced?

B: I think that there's a lot of room, just in the design of it for unreliablility, but at the same time, given an evaluation that's something you have to constantly live with. But I think you made a strong effort with your interrater reliablility scores and by letting us see the items. It makes it fairly easy to assess it, but I think we have to still live with the fact that some interviews may have taken a different tone than others did based on how well they knew us and how well you know of their programs and all that. So I think that there's room for unreliability, but I think that you just made a real good effort at removing as much of that as you can from it.

A: I would agree with that.

SPECIFIC QUESTION: Was there evidence that the data were collected and analyzed systematically?

B: I'd say "yes."

A: Yes.

E: Was there any area where error may have entered into the process?

B: I'd say the cognitive tests just by the constraints that we had for them. They could have easily been done in groups or with the information right there in front of them. Looking at their answers I don't think they did that, but I think that was certainly a possibility.

A: I'd agree.

SPECIFIC QUESTION: Did it appear that the quantitative data were appropriately and systematically analyzed?

A: From what I saw, I would say yes.

B: Yes, I would agree. I think that you made a good effort at analyzing it.

SPECIFIC QUESTION: Did it appear that the qualitative data were appropriately and systematically analyzed?

A: Yes, as much as you can subject any of that data to analysis.

B: And also it tends to get a little volumninous at times. I think that's where I got inundated with information.

SPECIFIC QUESTION: Were the conclusions presented in the evaluation report supported by the data?

A: Yes.

B: I'd agree.

RESEARCH QUESTION #5: Were the methods/instruments used within the evaluation framework ethical in dealing with people and organizations?

SPECIFIC QUESTION: Was the evaluation report open, direct, and honest in its disclosure of pertinent findings, including the limitations of the evaluation?

A: Yes.

SPECIFIC QUESTION: Was the evaluation designed and conducted so that the rights and welfare of the human subjects were respected and protected?

A: Yes, I think that you were up front with them at all times and there were no hidden agenda. You did not tell them they were being audiotaped to improve the quality of responses and to make them more relaxed. Also to improve the accuracy of the responses you recorded.

Additional questions were asked that did not relate specifically to any of the research questions. These questions did relate to the question of how well the evaluation framework had functioned when it was applied to the September session of the FMFDP.

E: What would you do if you were to do it again?

A: I think it is a job for someone other than the director or assistant director. In terms of a report I would do the End-of-Week Evaluations, I would do the cognitive pre-, post-, delayed posttests, although I would want that to be a different type of cognitive measure. I would do the videotape presentations and ratings. I would do the final debriefing. I guess it's the interviews that I have some question about in terms of the cost, the time, and I think we can get the data from other sources.

B: Yes, I think I would agree with that because we are getting an awful lot of information directly from the fellows. They have input at the debriefing, with the End-of-Week Evaluations, and it seems to me that the interviews are an expensive way to add to that.

E: How about interviewing the supervisors?

A: I guess in my mind that other than for public relations payoff that

I don't see that that contributed a lot in terms of information. It gave us a lot of satisfaction indices, but in terms of providing us with information that we could use to improve and revise the program, I didn't see a lot there.

B: I guess I'm a little more split in that sometimes I think the PR factor actually does have its own positive impact on the evaluaton and as you get closer to understanding the types of settings under which they're working, the kinds of responsibilities that they are initiating rather than just being given. I can't remember if it was read or it I was picking it up from you along the way, but I was learning things that I might not have learned through any other source.

E: What was not done that should be done?

B: You know what I would do? I would in some way incorporate the two visits we make to a setting, the one for the pre- and the one during the fellowship, somehow incorporate that into a data collection effort. I don't have specifics, but I think I would use that somehow to collect data.

E: Which evaluation source provided the information of most value to you?

B: The participants.

A: The fellows.

E: Which evaluation method provided the information of most value to you?

A: For me, the ratings of the actual performance, the presentations. It depends on which question, let me qualify it that way. The basic question that I was interested in was "To what degree is what we're doing transferrable?" In other words, what can they now put into operation and that data helped me a great deal. In terms of overall acceptablility, the End-of-Week Evaluations are what I look at.

B: Almost exactly. There's not a single instrument that I would have my confidence in, but if I had to pick two working together, one where we directly observe their performance and one where they give us information on how well we did on our presentations, the two together would be the best guess.

E: Which type of data would you rely on?

B: Behavioral.

A: Performance data.

E: Which would you least rely on?

B: Cognitive.

A: The cognitive and the supervisors' report.

E: What was the overall strength of the evaluation?

A: The strength for me is that it did make an attempt to collect data about both cognitive and affective outcomes and it used a variety of different sources of data, different data collection methods, and that there was some redundancy and I think that's good.

B: I think the strength of it was that you managed to hone in on what I think are the three best types of information. The weakness I felt was that in an attempt to be comprehensive we might have taken in too much information. I think maybe we had too many measures under each of those different types. That's what I've learned from it anyway, we didn't know that going in.

E: What are you doing this time?

B: It may be a little early to tell. We haven't implemented any of the additional things that you've done yet, primarily our questions were answered, the big ones. I don't think we've made any decisions if we're gonna do any of the follow-up interviews yet, but it may affect how we gather information when we go out to do the site visits.

A: We certainly videotaped them. No thought of doing any analysis. It's not because we don't feel it's important, like I said earlier this is the fifth year of the program and we have been doing the same thing for three years and your evaluation came in the fourth year and provided confirmatory data. In other words, we got confirmed that what we were doing was right.

B: Another thing along with that. I think what we discovered was that some of our initial assessment of their performance was substantiated with fairly rigorous ratings of those. I don't think we were that far off.

A: It's not like they weren't rated before. We sat down and we had our own criteria that we judged them against and that we gave them feedback on. What the rating process did was just formalize that more.

E: Anything else?

B: I was real happy with it.

A: Nothing other than I think you did a real nice job.