DEVELOPMENT AND VALIDATION OF A MODEL EXPLICATING THE FORMATIVE EVALUATION PROCESS FOR MULTI-MEDIA SELF-INSTRUCTIONAL LEARNING SYSTEMS

Thesis for the Degree of Ph. D.

MICHIGAN STATE UNIVERSITY

ALLAN JOSEPH ABEDOR

1971

This is to certify that the

thesis entitled

DEVELOPMENT AND VALIDATION OF A MODEL EXPLICATING THE FORMATIVE EVALUATION PROCESS FOR MULTI-MEDIA SELF-INSTRUCTIONAL LEARNING SYSTEMS

presented by

ALLAN JOSEPH ABEDOR

has been accepted towards fulfillment
of the requirements for

Ph. D. degree in Secondary Education

Major professor

Date August 13, 1971

O-7639

ABSTRACT

DEVELOPMENT AND VALIDATION OF A MODEL
EXPLICATING THE FORMATIVE EVALUATION PROCESS FOR MULTI-MEDIA
SELF-INSTRUCTIONAL LEARNING SYSTEMS

By

Allan Joseph Abedor

Tryout and revision are steps considered by many to be essential to the development of an instructional system. Virtually all theoretic models of instructional system development include tryout and revision as an integral part of the process. However, the formative evaluation procedures included in such models are either too general to be useful, or when specific, seem applicable to simple textual programmed instruction.

New tryout and revision procedures are needed to operationally apply the principles of formative evaluation to instructional systems of increased complexity and scope. The purpose of this study was, therefore, to develop and validate (field test) a flowchart or analog model prescribing specific formative evaluation procedures for tryout and revision of prototype multi-media self-instructional learning systems.

The initial (MK I) model was developed from a review of the literature on formative evaluation. This model addressed three main methodological issues: (1) how to identify major discrepancies in prototype multi-media lessons by data collection; (2) how to analyze these data and develop revision hypotheses; and (3) how to design, integrate, and evaluate revisions. The MK I model stipulated an elaborate three-stage process, including technical review, tutorial tryouts, and large group tryouts.

Validation of the MK I began by having its procedures assessed by means of interviews with seven faculty members who had previously developed (and revised) multi-media lessons. Data from these interviews clearly showed that the MK I procedures were far too complex and time consuming for practitioners to use. Therefore, an MK II version was developed which simplified procedures throughout and introduced a small group (N=12) tryout and debriefing procedure as the main method of identifying instructional problems and developing revisions.

This technique required nine to twelve volunteer students of varying ability to individually interact with prototype lesson materials. During student use of the prototype, the lesson author personally answered questions in a tutorial fashion. After completion of the lesson, students were given a 15-minute recess so the lesson post-test and attitudinal survey could be scored. Items which indicated that 30% or more of the group were having problems were tallied and became the agenda for the debriefing to follow. During the debriefing, which was conducted by the lesson author, students were encouraged to freely discuss any and all problems they encountered--and to provide solutions to these problems if possible. The identification of prototype lesson problem areas and development of revision hypotheses thus became an author/student group responsibility.

Validation of the MK II procedures were conducted in five field experiments conducted with three Michigan State University faculty, Fall term, 1970. The purpose of the experimental comparisons was to determine, insofar as possible, the overall validity, feasibility, and effectiveness of the MK II model in facilitating tryout and revision of prototype multi-media lessons.

Faculty member A had developed three prototype multi-media lessons, designated A_1, A_2, and A_3. Faculty member B and C had developed one lesson each, designated B_1 and C_1. Each field experiment consisted of the lesson author applying the MK II procedures to tryout and revision of his prototype lesson. In each field trial, the experimenter (E) performed technical assessment of prototype instructional stimuli, after which the materials were tried out with the first student group. Following the first tryout and debriefing, revisions suggested by the students were incorporated into revised versions.

As revised lessons were completed, a second iteration of student tryouts was initiated. The purpose of the second tryout was twofold: (1) to compare the revised version with its prototype counterpart to determine the effect of the revisions on measures of student attitude and achievement; and (2) to gather additional feedback for further revisions. On two trials (A_3 and C_1), however, after the first student tryout the authors concerned felt that the initial prototype was sufficiently effective and did not warrant revision. Hence, in these two cases, an experimental comparison between prototype and revised versions was not possible.

In the three trials in which experimental comparisons were conducted, simple statistical tests were used to compare four dependent measures: (1) student achievement on the post-test, (2) gain score, (3) percentage of students achieving criterion, and (4) student attitudes. In two field trials (A_1 and B_1), significant differences were obtained ($P < .01$) favoring the revised version on all four dependent measures. In the third field trial (A_2), a significant difference ($P < .05$) favoring the revised version was obtained on the post-test only.

It was concluded that: (1) the MK II model was valid, in that authors were able to identify and remediate major instructional problems through use of MK II procedures; (2) the MK II was feasible, in that two out of three authors were willing and able to use MK II procedures; and (3) the MK II model was effective, in that statistically significant differences favoring the revised versions were obtained on nine out of twelve dependent measures in the three separate field trials.

The MK II model provides an operational framework within which instructional development personnel can train or consult with faculty regarding formative evaluation of mediated self-instructional systems. Whether the model can be generalized to other types of instructional systems is a question yet to be answered.

The MK II procedures are developed at two levels of detail. The "mini" MK II is a simplified version used for orientation purposes. The "maxi" MK II provides the detailed procedures needed by an instructional development specialist.

The general principles of the model are as follows: (1) use a carefully developed prototype to provide a common instructional experience for a group of volunteer students of varying abilities; (2) collect data by means of learning and attitudinal measures after the common experience; (3) identify, discuss, and propose solutions to major problems by means of a group debriefing conducted by the author; (4) consult with "experts" on data interpretation; and (5) revise the instructional unit and recycle as necessary.

DEVELOPMENT AND VALIDATION OF A MODEL EXPLICATING THE FORMATIVE EVALUATION PROCESS FOR MULTI-MEDIA SELF-INSTRUCTIONAL LEARNING SYSTEMS

By

Allan Joseph Abedor

A THESIS

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

DOCTOR OF PHILOSOPHY

College of Education

1971

DEDICATION

This thesis is dedicated

to my Mother and Father.

ACKNOWLEDGMENTS

The writer wishes to express his appreciation to the many persons who have contributed to the design, development, and execution of this thesis.

Particular thanks are expressed to Dr. Paul W. F. Witt for his encouragement and counsel as chairman of the guidance committee and to Drs. Robert H. Davis and Norman T. Bell for their inspiration and guidance during critical phases of this thesis. Thanks are also expressed to Drs. Kent L. Gustafson and James R. Nord for their insightful suggestions.

Special thanks go to Dr. Harold A. Henneman, Dr. Howard H. Hagerman, Mr. Thomas W. Burt, and the students in their courses without whose cooperation this thesis would not have been possible.

Gratitude is also expressed to Drs. David K. Berlo, Randall P. Harrison, Lawrence T. Alexander, and Stephen L. Yelon who provided conceptual guidance during the formative stages of this thesis.

Deepest appreciation goes to my wife Betty, my daughter Carolyn, and my son John for their patience and forbearance in accepting the absence of their husband and father during much of the time the thesis was in progress. To them, I can only offer myself and my gratitude for the constant love, encouragement, and devotion they have provided.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

BACKGROUND

Tryout and revision are steps considered by many to be essential to development of an instructional system. Virtually all theoretic models of instructional system development include tryout and revision as an integral part of the process. For example, models developed by Barson (1965), Paulson (1969), Hamreus (1968), Briggs (1970) and Smith (1966), take the form of a flowchart describing a programmatic sequence of activities of which approximately the last one-third is devoted to tryout and revision.

Tryout and revision have long been recognized by writers in the field of programmed instruction as essential components of the program development process. According to these authors, programs should be tried out and revised until they meet some predetermined standard of student performance. Susan Markle (1967) cites the principle of "developmental testing" (her term for tryout and revision) as one of the major factors differentiating programmed instruction from conventional instruction.

There is some evidence that the principle of tryout and revision has been attempted with various types of instructional systems. Gropper, Lumsdaine and Shipman (1961) demonstrated increased student recall and retention after applying the tryout and revision process to conventional television lessons. D. Markle (1967) developed a first aid training course using films, texts, and practice based largely on empirical tryout

and revision. It would seem, therefore, in light of the emphasis given this topic in programmed instruction and instructional system development that the need for empirical tryout and subsequent revision would be well understood today.

Nevertheless, in a recent Review of Educational Research, Popham (1970) observes:

> From an inspection of the research related to curriculum materials during the past several years, one is impressed by several deficiencies. First, studies of the revision process to improve the quality of curriculum materials have not been clearly demonstrated. Certainly the manner in which revisions can be made most efficiently has not been carefully treated (emphasis added) (p. 335).

Later, in the same review, Popham quotes Lumsdaine as saying, "There was little research which demonstrated that revision based on empirical tests, as opposed to skilled editorial revision, produced better learner achievement (p. 331)."

There are two points to be stressed here. First is the paradox wherein many writers in programmed instruction, instructional technology, and instructional system development strongly advocate the principle of tryout and revision. Yet on the other hand, educational researchers seem to have ignored the topic. Perhaps it was felt that the principle was so self-evident that little corroborative research was needed.

The second and more important point is that the few research studies and theoretic papers which address this question usually deal only with tryout and revision of a simple instructional system of the size and complexity, for example, of a single (usually short) programmed text. The techniques and procedures used in tryout and revision of "pure"

programmed textual materials (such as error rate, response time, frame analysis, criterion frames, etc.) seem inappropriate or irrelevant for a lecture, a laboratory, a multi-media (slide-tape) presentation, or other instructional modes commonly employed together in a single instructional system.

Therefore, a research question of importance to the instructional development specialist is: What specific methods are appropriate for tryout and revision of complex, multi-component, instructional systems? In other words, how ought the principle of tryout and revision be implemented in developing an instructional system having several components such as lecture, laboratory, small group discussion, and multi-media self-instructional units? A more fundamental question is: How can instructional system designers utilize systematic feedback from students or others in the design process?

This question has several aspects. First, the available theoretic models of instructional system development are written at a very general level. Most of these models provide a "what-to-do" orientation, but not "how-to-do-it" detailed information. The few models which do try to provide specific "how to" information invariably recommend procedures drawn directly from simple programmed instruction texts and these do not appear generalizable to other modes of instruction. How, for example, does one compute an error rate or frame analysis of a lecture, laboratory, recitation or film presentation?

Another aspect of this problem is that available procedures for tryout and revision focus almost exclusively on identification of general problems, with little guidance on specific remediation. Obviously,

problem identification is critical, but general identification per se does not necessarily indicate what the specific solution, or range of solutions ought to be.

It is the central assumption of this study that new methods of tryout and revision must be developed for complex instructional systems, and their components. Further, that tryout and revision methods must go beyond problem identification and develop viable techniques for remediating deficiencies and improving the product. At present the available guidance on tryout and revision is either too general to be useful, or when specific--directed towards simple textual programmed instruction. What is needed is an extension of previous research to develop detailed tryout and revision procedures which are adaptable to systems of increased complexity and scope.

Purpose of the Study

This study attempted to explicate the tryout and revision aspect of the instructional system development process. This explication included the development and validation (field test) of a flowchart model and a set of heuristics for applying the model to the tryout and revision of multi-media self-instructional systems. Such multi-media systems represent a far greater level of stimulus complexity than textual programmed instruction, so new procedures were developed for both problem identification and remediation. In sum, the purpose of this study was to develop techniques which enable systematic feedback from students and/or others to be used as an integral part of the development process used in the creation of multi-media self-instructional systems.

Assumptions

It was assumed that the model developed in this study represents an expansion of a part of the larger process of instructional system development (ISD) and should therefore be compatible with existing models of this process. The present day ISD models all include a tryout and revision phase, thus selection of an ISD model within which to embed the expanded tryout and revision model was based on the experimenter's previous familiarity rather than any functional differences.

The ISD process model within which the tryout and revision model developed in this study was assumed to operate is the Hamreus (1968) "maxi" version shown in Figure 1. It is important to note that the flowchart model developed in this study attempted to specifically explicate steps 16 through 22 in the Hamreus model.

Limitations of the Study

The flowchart models of tryout and revision procedures developed in this study were designed for and validated with a single type of instructional system; namely, a multi-media self-instructional presentation. This type of instructional system was selected because: (1) mediated instructional stimuli may be replicated exactly, providing greater experimental control than many other instructional modes; and (2) increasing numbers of university and community college courses use multi-media self-instructional units to accomplish a large part of the instructional function.

Selection of this class of instructional subsystem is not to be construed as an evasion of the problems of tryout and revision of less replicable or less controllable instructional subsystems (lectures,

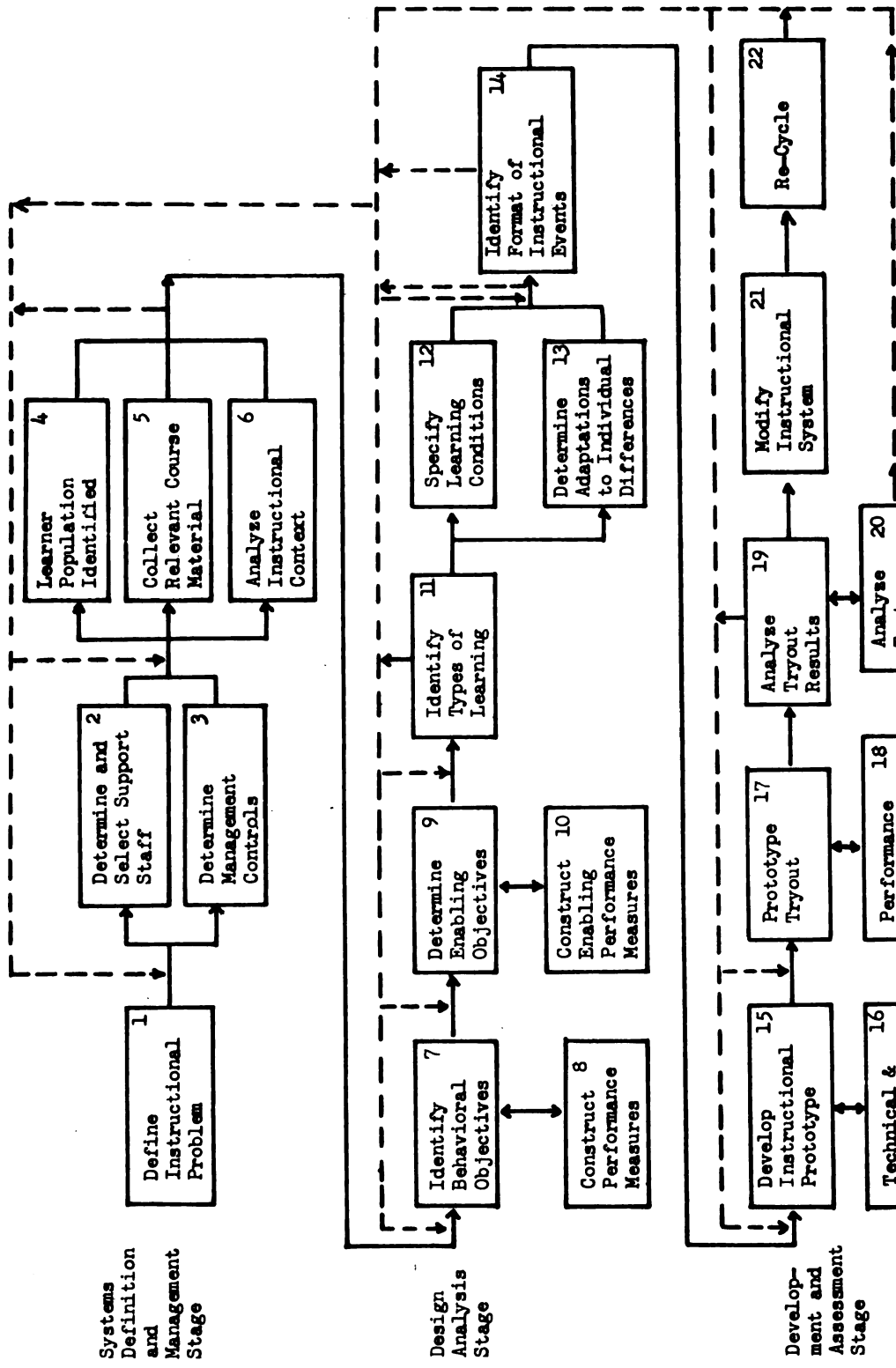


Figure 1. Flow diagram showing the specific steps of the Systems Approach in developing instructional systems

laboratories, recitations, etc.) or evasion of the problems of tryout and revision of the "course" as a total system. On the contrary, it was felt that the tryout and revision model developed for mediated self-instructional lessons may be generalized to the more emergent, spontaneous, non-mediated subsystems as well. However, the generalizability of the model was not specifically investigated in this study.

Another limitation of this study related to the difficulty in differentiating between unique contributions of personnel using the model versus the contribution of the model per se. It was assumed that the lesson author, the experimenter, and the students made unique individual as well as interactive contributions to the revision process. These unique contributions were not necessarily reflected in the formalized model or methodology. Thus, it became very difficult to assess what part of the differences between prototype lessons and revised versions were due to use of the model/method or due to the unique contribution of the personnel involved.

In some cases, differential contribution of method and personnel variables can be identified by sophisticated experimental design. In this study, however, a new model was conceptualized; consequently, it was not feasible to hypothesize specific relationships between methodological and/or personnel variables. Instead it was assumed that the first step in extending the tryout and revision process to instructional systems of greater complexity than simple programmed texts was to develop and describe a workable, viable model.

Organization of the Thesis

In the balance of the present chapter, the organization of the thesis, its major objectives and methodology are described. Limitations and assumptions are stipulated and key terms defined.

In Chapter II, literature relevant to tryout and revision are reviewed and a preliminary flowchart model (MK I version) developed.

In Chapter III, the results of interviews with seven experienced multi-media lesson designers are presented along with a rationale for revision of the MK I model and development of the MK II version.

In Chapter IV, the descriptive and experimental methodology for five field tests of the MK II version are outlined.

In Chapter V, the results of the five field tests are described and the experimental data reported and analyzed.

Finally, in Chapter VI, the major findings of the study are summarized, conclusions drawn and recommendations for further research provided.

Definition of Terms

Formative Evaluation

Formative evaluation involves the tryout and revision of new instructional units in an effort to improve quality prior to large scale use with students. As used in this study, the term is synonymous with "tryout and revision" or "developmental testing." Generally, formative evaluation is the process by which information is obtained and used by a decision maker to identify problems and revise instruction to the point where it is ready to be used with substantial numbers of students. The

decision maker of interest in formative evaluation is the developer of the new instructional system.

In this study, the process of formative evaluation was conceptualized as having three components: (1) identification of instructional deficiencies through data collection; (2) analysis of these problems leading to revision hypotheses; and (3) design, integration, and evaluation of revisions.

Scriven (1967) defines formative evaluation as "outcome evaluation of an intermediate stage during development of the teaching instrument . . . to discover the deficiencies and successes in intermediate versions of new curriculum (p. 51)." Anderson (1969) emphasizes that "the purpose of pilot tests is formative evaluation, to locate weaknesses in student understanding or performance so that editors, writers, or teachers can revise and presumably improve instructional materials and procedures (p. 5)."

Summative Evaluation

Summative evaluation is the process of describing the effects of such fully developed units of instruction (Paulson, 1969). The decision maker of interest in summative evaluation is the consumer or user of the instructional system rather than the developer. Both formative and summative evaluation are emphasized in ISD models and both reflect the basic principle that any system requires feedback to achieve its objectives (Wiener, 1954).

SLATE

SLATE is an acronym for Structured Learning and Teaching Environment (Davis, 1968). Typically, a SLATE involves a single student in a

carrel interacting with multiple instructional stimuli in the form of slides, tape, film, models, specimens, and a workbook. The learning experience is "structured" in that objectives are predetermined and students' responses are designed to facilitate achievement of these objectives. A SLATE is, therefore, a multi-media self-instructional learning system.

Flowchart Model

A flowchart model is a graphic analog showing the total structure, organization, and interrelationships of a process, event, or other phenomenon. In the present study, flowchart symbols represented ideas, information flow, and human action with narrative explanation being provided for each symbol. The LOGOS symbol system (Language for Optimizing Graphically Ordered Systems) developed by Silvern (1969) is used in this study.

Author

As used in this study, the term "author" refers to a faculty member who has developed one or more multi-media self-instructional lessons.

Prototype

Prototype refers to a complete, but untried version of a self-instructional multi-media lesson. In other words, all instructional stimuli are finished, but student feedback on the efficacy of these stimuli has not been obtained.

Debriefing

Debriefing refers to a formalized procedure where through face-to-face interaction the prototype lesson author obtains information from students on lesson deficiencies and how to remediate these deficiencies.

Methodology of the Study

The present study used an exploratory approach to develop and validate the new model of formative evaluation. The exploratory approach is described by Kaplan (1964) as follows:

> It (an exploratory study) is frankly intended just to see what would happen if . . . . Often it is associated with a new technique, which is tried on a wide variety of problems and subject-matters until the most promising sorts of applications become apparent . . . . Or, it may be conducted according to a trial and error pattern to exhaust some set of possibilities. In general, an exploratory experiment invites serendipity, the chance discovery; it is part of what we do to deserve being lucky (p. 149).

In the present study, it was intended to see what happens when the "new technique" (the model) was tried in a series of five different "field experiments" (Kerlinger, 1964) involving three different academic subjects.

Since previous research (Baker, 1970; Silberman & Coulson, 1965) has clearly shown simple programmed texts to be significantly improved through use of certain tryout and revision procedures, the functional outcome of this study was to tell us something new only about a certain way of achieving such results in other cases; e.g., in an instructional system of increased complexity, such as a multi-media self-instructional lesson.

Having this exploratory orientation and methodological purpose the study organizes naturally into two phases: (1) design of the theoretic model, and (2) exploratory field test (validation) of the model. Each phase is fully described in a later chapter of the thesis. For orientation purposes, however, the major objectives and methodology of each phase are described next.

Theoretic Phase

Objectives.--The major objective of the first phase was to develop a flowchart model showing a sequence of tasks, decision rules, criteria, and implementing methodology for empirical tryout and revision of a prototype multi-media self-instructional lesson. The criteria for assessing the utility of this model were threefold.

1. Validity.--The model was considered valid if (a) through its use the prototype lesson author was able to distinguish those sequences of instruction which were unsatisfactory, and (b) if the model predicted revision alternatives which remediated the unsatisfactory instructional sequences.

2. Feasibility.--The model was considered feasible if fewer than 20 students were required for its use, and if faculty were willing and able to use it in the field situation.

3. Effectiveness.--The model was considered effective if comparative measures of student achievement and/or attitude between prototype and revised versions showed statistically significant differences in favor of the revised versions in 75% of the field experiments.

Methodology.--The model was developed from two primary data sources: (1) review of pertinent literature, and (2) interviews with a selected sample of university faculty or research and development personnel who have personally developed multi-media self-instructional lessons. The purpose of these interviews was to: (a) get faculty reactions to the model, and (b) assess and integrate (if possible) the tryout revision procedures actually used by practitioners.

The interview data were summarized to highlight the procedures and/or recommendations common across respondents. These data were then used to modify the flowchart model developed purely from the review of research and theoretic literature. The product of the first phase therefore was a synthesis of research, theoretic and practitioner data stated as a "first draft" flowchart model of tryout and revision procedures.

Exploratory Field Test Phase

Objectives.--This phase was somewhat unique in that two distinctly different types of objectives were being sought, e.g., product and process. The product type of objectives relate to experimental comparisons between two lessons to determine which is superior. In this study, revised lessons were experimentally compared to their original (unrevised) counterparts. The intervening or independent variable was the use of the tryout and revision model, and the dependent variables were measures of student achievement and attitudes. By empirically comparing a given lesson before and after tryout and revision, tentative conclusions regarding the validity, feasibility and effectiveness of the revision techniques could be drawn.

A second, and possibly more important class of objectives centered on understanding and describing the process through which the deficiencies in any given instructional system were recognized and remediated.

The model per se, was simply a conceptual tool designed to influence a series of complex interactions between human beings; e.g., faculty, students, and consultants. It was these interactions which resulted in modifications to a given instructional system. Therefore, it was important to assess the nature of those interactions to determine what factors, besides

the model, were influential in developing the revised instructional system. In short, it was important to describe and analyze the interactive process of conducting tryout and revision so that major variables could be identified and the procedures guiding the process modified to take account of these variables.

To summarize, the overall objective of the field test phase was to generate and analyze empirical data from which tentative conclusions could be drawn regarding the efficacy of the model. In addition, this phase provided descriptive data for: (1) identification of variables in the formative evaluation and revision process of a given instructional development system, (2) recommendations for procedural modifications and/or alternatives which take account of such variables, (3) recommendations for further research.

Methodology.--In light of the two types of objectives just described, the methodology used in this phase combined experimental and descriptive techniques. That is, in addition to conducting several field experiments, the experimenter (E) described the process through which the experimental treatments generated.

A total of five (5) field experiments were conducted in which measures of student attitude and achievement on a prototype (unrevised) lesson were compared to identical measures of achievement and attitude obtained on a revised version. A field experiment is defined by Kerlinger (1964) as follows:

> A field experiment is a research study in a realistic situation in which one or more independent variables are manipulated by the experimenter under as carefully controlled conditions as the situation will permit (p. 382).

A field setting was chosen because it was desired that the relationship between use of the model and improved student performance be demonstrated in the actual using environment. In this way, unforeseen or contaminating variables are identified and accounted for in revisions to the model.

Each experiment employed the classic pre-post control group design (Campbell & Stanley, 1963) in which a sample of volunteer students (Ss) were stratified and randomly assigned to experimental and control groups. Control groups received the prototype (unrevised) SLATE, while experimental groups received the SLATE revised in accordance with the model of formative evaluation. The independent variable in each case was the model plus unique contributions by the users of the model. It was hypothesized that revisions resulting from student feedback should produce gains on achievement and attitudinal measures in the experimental groups. T tests were used to determine statistical significance between experimental and control groups.

Unlike more conventional experimental/control group comparison research, the experimental treatments in this study were not designed a priori in accordance with some set of theoretic principles. Instead, feedback from students who had utilized the original prototype lessons provided three types of data which were used by the lesson author and experimenter to develop revisions. These types of data were: (1) measures of student achievement, (2) measures of student attitudes (both 1 and 2 were collected during the lesson via specific instruments), and (3) experiential data generated during an author/student debriefing following the lesson. Each experimental treatment evolved by means of

several student/author/experimenter interactions prescribed by the theoretic model.

The total process of trying out a prototype lesson, revising it, and testing out the revised version was replicated five times. Three individual Michigan State University faculty in different disciplines were the prototype lesson authors. Two of these faculty members revised one lesson each, while a third faculty member revised three separate lessons. The selection of five replications and three academic disciplines was based on (1) availability of faculty with unrevised prototypes, and (2) the need to provide a sufficient number of cases from several disciplines from which to identify critical variables in the process and determine the efficacy of the model.

The series of five field experiments were organized chronologically so that the experimenter was able to assist individual faculty as required, as well as observe and document the entire revision process for each experiment.

After obtaining verbal commitment from the faculty indicating their interest and willingness to invest a substantial amount of time in revision activities, the experimenter began formal observations and provided assistance to each faculty in applying the model to their prototype lessons.

The assistance given each faculty was of two types, logistical and conceptual. The logistical help consisted of making the physical and administrative arrangements necessary for the field experiment, e.g., to tryout the lesson on a selected sample of students.

The conceptual assistance given the faculty was essentially to explain the model, justify its theoretic orientation and attendant methodology, and provide guidance as required in performing the tasks specified in the model. In many ways this was a tutorial training function, since in its initial stages of development the model was not self-explanatory, although this was the long-range goal. After experimenter guided utilization on several lessons, it is possible that faculty may understand the model sufficiently to apply it independent of the experimenter, but this contingency was not specifically tested in this study.

Potential Payoff From This Line of Research

Greater sophistication in the tryout and revision phase of instructional systems development can lead first of all to fewer and fewer revisions and improved learning from students. More importantly, it might lead to the formulation and assessment of principles which, if incorporated in initial preparation of instructional units, could lead to units requiring minimal revision. In other words, a highly developed instructional technology may always require empirical tryout, given the individual differences in human learners. But, the more sophisticated the body of procedures, techniques, and principles used in initial design, the better the initial preparation of new instruction can be. In short, greater sophistication in tryout and revision can lead to principles which may improve the process of initial design.

CHAPTER II

REVIEW OF THE LITERATURE LEADING TO DEVELOPMENT OF A PRELIMINARY (MK I) MODEL OF FORMATIVE EVALUATION

The review of formative evaluation research presented in this chapter addresses three methodological issues: (1) how to identify problems in prototype instructional units, (2) how to analyze such problems and develop revision hypotheses, and (3) how to design and integrate revisions.

The review is organized as follows. First, the assumptions underlying the research reviewed and derivation of the MK I model are stipulated followed by specific questions relating to problem identification, problem analysis, and problem remediation. These questions form the basis for analyzing the research of nine selected authors in the field of formative evaluation. Next, the research of the selected authors is described followed by an analysis to determine the given author's specific approach to the methodological questions stipulated earlier. After the last research study has been presented and analyzed, several conclusions with respect to the three methodological issues are drawn and a preliminary MK I model presented. The model thus represents an integration of the literature reviewed.

Assumptions Underlying Development of the MK I Model

The selection of literature for review and the conclusions reached thereafter were based largely on the assumptions and definitions

stipulated in Chapter I. Briefly, the most critical of these were: (1) formative evaluation is basically data collection and use of information by a decision maker to revise deficient instructional sequences; (2) the critical decision maker in this study is the author/developer of a self-instructional multi-media lesson; and (3) the primary source of information relative to identification of instructional deficiencies should be students for whom the prototype lesson was intended. A secondary source of information may be "experts."

Specific Questions Used to Focus the Review of the Literature

The following questions were used in analyzing each of the research studies reviewed.

ISSUE	SPECIFIC QUESTIONS
HOW TO IDENTIFY PROBLEMS IN PROTOTYPE INSTRUCTIONAL UNITS	1. What types of data were collected? 2. What types of instruments were used and how were they developed? 3. What sampling procedures were used? 4. What administrative procedures were used?
HOW TO ANALYZE SUCH PROBLEMS AND DEVELOP REVISION HYPOTHESES	5. How were data reduced and interpreted? 6. How were priorities assigned among instructional deficiencies? 7. How were revision hypotheses developed?
HOW TO DESIGN AND INTEGRATE REVISIONS	8. How were revisions designed, integrated, and evaluated?

Essentially, the review of the literature attempted to fill in each cell of the matrix shown in Table 1. The completed matrix (Table 3)

Table 1.--Matrix Showing Organization of the Review of the Literature

Issue	Specific Questions	Authors Reviewed		
		A_1	A_2	A_3
Problem Identification	What Data Types			
	What Type Instruments			
	What Sampling Procedures			
	What Administrative Procedures			
Problem Analysis & Revision Hypothesis Development	How Data Reduced & Interpreted			
	How Priorities Assigned			
	How Revision Hypotheses Developed			
Revision Design: Integration & Evaluation	How Revisions Designed & Integrated			
	How Evaluated			

thus functions as a transition device to summarize the data which was incorporated into the MK I flowchart model.

Review of Research by Individual Authors in Formative Evaluation

Generally, the literature on formative evaluation presents three different approaches or evaluation design strategies (Paulson, 1969): the "tutorial" or single student feedback approach, the "large group" or multi-student feedback approach, and some iterative combination of the first two. Each approach has unique advantages and disadvantages.

The Tutorial Approach

The writings of B. F. Skinner (1954, 1958) introduced the fundamental concepts of linear programmed instruction and teaching machines. Skinner's laboratory-like technology emphasized study of the individual organism and precise control of behavior to be learned by manipulating the consequences of behavior.

Following Skinner's lead, many writers in programmed instruction have advocated that the optimal unit of analysis for development of programs was a single student. For example, Gilbert (1960) suggested twelve rules for programming a specific subject matter. Rule five is:

> Get yourself one student. I repeat, one student. You are about to perform an experiment in which you are permitted no degrees of freedom--that is, if the word "self" in "self-instruction" can be taken seriously. Once you have discovered an efficient program for one student, you will have described the gross anatomy of the most generally useful program (p. 479).

Some empirical evidence in support of the tutorial approach and a single student as the unit of analysis was provided by Robeck (1965) and Silberman and Coulson (1965).

Research by Robeck.--Robeck demonstrated empirically that observation of a single student can significantly improve a first draft program. Using a short (50 frame) prototype PI text on "English Money," he incorporated revisions based on test item errors and verbal responses of a single "bright" sixth-grade student to produce a second draft. This draft was further revised on the basis of feedback from a second individual student. The three drafts were then tested on three matched groups of students. The performance on the two revised versions was significantly better than on the original draft ($P < .05$ for the second and $P < .01$ for the third version) although student performance on the third version was not significantly improved over the second version. While the revisions depended on the ingenuity of the tryout editor-tutor as well as verbal data from single bright students, Robeck did demonstrate the feasibility of a single solitary student as the total sample for formative evaluation.

Unfortunately, the study was not clear as to the sampling procedures (how the "bright" students were selected) or the procedures used during the tutorial interaction to identify discrepancies. Moreover, the study was not clear as to how the test performance and error rate data was integrated with verbal responses of the students to identify causal factors, so revision hypotheses could be developed. Further, Robeck provided no information as to how the revisions were designed and integrated into the original program. He reported, however, that evaluation of the revisions was obtained through a suitable experimental/control group design.

The implication of Robeck's research for the present study was that achievement and interview data from a single student could be

used to complement one another in development of an improved programmed text.

Research by Silberman and Coulson.--A far more extensive and sophisticated set of experiments was conducted by Silberman and Coulson (1965). In this series of exploratory studies, a technique called "tutorial engineering" was developed in which an experimenter served as a tutor while presenting the program (PI text) to one child at a time. The experimenter stopped the presentation and provided tutorial assistance whenever the student exhibited difficulty (cues were verbal, "I don't understand"; non-verbal, a puzzled look; and test item errors). Tutorial assistance was ad hoc, but records were kept of student difficulties and tutorial procedures which seemed most beneficial. Similar tutorial assistance needed by more than two students was incorporated into the programmed text as revisions.

When the experimenter-tutor felt that the program had been revised sufficiently (sufficiently was not operationally defined but was a subjective decision) a comparison of the original and revised program was made. If the revised version proved to be both statistically and practically (not too much longer than the original) superior to the original, the tutorial sessions were ended; if not, the experimenter conducted several more cycles of tutorial engineering.

A total of four programs representing verbal and quantitative skills was developed in this manner (first grade reading, first grade arithmetic, junior high Spanish, and senior high geometry). After all four programs were significantly improved, the data collected during the tutorial sessions and the student responses to different versions of the

programs were analyzed for consistencies and patterns. This analysis produced three hypotheses about major instructional problems which were common to all four programs. These hypotheses were termed the "gap," "irrelevancy," and "mastery" hypotheses. The "gap" hypothesis refers to the necessity for explicit inclusion in the program of information relevant to each criterion test item. The "irrelevancy" hypothesis refers to the desirability of eliminating material which is unrelated to criterion test items. Finally, the "mastery" hypothesis refers to the requirement that the student not be permitted to move on to subsequent topics until he had "mastered" the present one.

Since these three hypothesis had been derived by analysis of revisions to four programs, an independent experiment was then conducted with a different set of PI texts and new students to test the hypotheses. The new experiment validated these findings by reversing the process. It was shown that when these designated improvements were taken out of effective programs, there was a corresponding performance decrement.

The importance of Silberman and Coulson's series of studies was that for the first time, the formative evaluation and revision process was formalized (the "tutorial engineering technique") and at the same time empirically proven to be successful. Moreover, generalizations were drawn leading to higher order revision principles: the gap, mastery, and irrelevancy hypotheses.

One implication of Silberman and Coulson's work for the present study was that discrepancies in prototype programs were identified by a combination of student errors on achievement tests and tutor observation of students' non-verbal behavior (such as frowns). A second implication

was that a number of different tutors were used indicating that non-verbal data was recognizable and provided important feedback in a number of cases.

The research was unclear on the administrative procedures used during the tryout sessions; hence it was difficult to separate the unique contribution of each tutor from the general procedure of tutoring a single student as required and recording all tutorial interaction for later inclusion in the lesson. Furthermore, the procedures used for selection and training of tutors and selection of students were unclear.

The research did indicate, however, that interpretation of data and development of remediation was a joint decision between tutor and project directors. If several students were tutored on the same problem, this information was usually added to the program increasing its length. It did not seem as though program objectives were subject to revision; rather the four programs were simply lengthened to permit students to achieve the objectives. An absolute performance criterion was not established, so revisions were apparently ended when the tutor felt that students could use the materials without further tutorial assistance.

While many of the "tutorial engineering" procedures were not described as precisely as one would like, it was clear that given a prototype PI text, a tutor, and perserverance, it appears feasible to successfully revise a program based on multiple tutorial sessions using single students to provide feedback.

Theoretic work by Horn.--The clearest example of the "tutorial approach" was provided by Horn (1966) who developed a self-instructional program called Developmental Testing. This programmed text was designed

to train evaluators and/or programmers in the tutorial approach to formative evaluation. Horn not only programmed the elements of his technique, but presented simulated problems in formative evaluation--and provided feedback on appropriate solutions.

Horn asserted that his approach could be used successfully with as many as four students simultaneously, although he advocated the use of one student of relatively high ability, one of average ability, and one of low ability, singly, and in that order. This notion of progression from high to low ability students was similar to procedures suggested by Scott and Yelon (1969).

Many of the procedures recommended by Horn were similar to those used by Silberman and Coulson, but Horn went far beyond earlier works in his explication of the administrative procedures or "ground rules" which should apply during a tutorial tryout and revision session. In order to provide the reader with a clear understanding of this methodology, Horn's "checklist" for tryout sessions and "principles for determining when to intervene in the tryout process" are shown in Appendix I.

While Horn's text is an important contribution to the literature on formative evaluation, it was unfortunate that no empirical data as to the success (or failure) of the procedures was included. On the other hand, Horn's work may be considered as an explication of the "tutorial engineering" approach which was empirically tested by Silberman and Coulson (1965).

Descriptive research by Dick.--While the tutorial approach has been advocated theoretically and has some empirical support, a study by Dick (1968) showed as one of its findings that non-professional

inexperienced program writers preferred to base revisions on data from a large sample (N=40 to 50) rather than from individual students.

In this study, Dick developed a method for integrating seven types of feedback for revising prototype programs (Appendix J). This method consisted of a set of seven decision rules which stipulated techniques to be used for data interpretaion and revision design. The task given to four non-professional, inexperienced program writers was to revise a previously used programmed text in calculus using the seven decision rules and the various types of data provided.

Each of the four revision programmers was given the original program plus seven types of data collected the previous year from four intact classes totaling eighty-five students. These data included: item analysis of post-tests, error rate, student comments, teacher comments, list of correct and incorrect answers for all test items, and page number where a specific test item was taught in the text.

During the revision process, Dick found that the revision authors were utilizing two primary data sources: error rate and teacher comments. If the error rate became excessive (which depended upon the individual writer's opinion) teacher and sometimes student comments were studied and revisions developed accordingly. Few revision authors used item analyses or tried to relate test item performance with particular frames in the program. Moreover, it was clear that none of the revision authors followed the suggested sequence of seven rules. These four authors reported that the end-of-lesson tests, which they had not constructed, were inadequate tests of the lesson objectives. Furthermore, the revision authors were interested in knowing more about the ability level of the students

who had made specific comments on segments of the program and wanted data on the student's overall impression of program continuity, readability, and difficulty.

With respect to questions of sampling and administrative procedures, Dick summarizes his findings as follows:

> It was of interest to the author to note that when the writers were given a hypothetical alternative of gaining information about the program by going through it personally with three or four students vs. gaining statistical data from 40 to 50 students, the latter procedure was much preferred. There seemed to be a greater number of students (which appears to provide greater generalizability) and an acknowledgement of difficulty of obtaining suitable guinea pig students (p. 101).

Dick failed to provide background data on the four revision authors, so it was difficult to extrapolate his findings to any subset of potential program writers. Nevertheless, some implications may be inferred for the present study. First, assuming inexperienced program writers (or SLATE authors) it is possible that large amounts of different types of data will overload the revision author's decision making capability in spite of decision rules provided. In short, certain types of data are likely to be ignored by inexperienced programmers, probably because they do not know what to do.

This leads to the second implication; that revision feedback should be restricted to a few critical types of data and/or consultative help provided during the data analysis phase of the revision process.

Discussion of the Tutorial Approach

The tutorial approach to formative evaluation means simply that a tutor--either the programmer or a qualified assistant--sits down with a student as he interacts with the prototype materials and carefully observes the student's response to each frame or step in the program. If

the student encounters difficulty, he describes the problem to the tutor who verbally provides the needed information and makes an on-the-spot revision to the specific frame(s) causing the problem. Several empirical studies have shown that given the proper conditions of cooperative students and skilled tutors, the method can produce improved instructional sequences. However, this procedure appears deceptively simple, for its success depends on interpersonal subtleties which are difficult to formalize into statable principles. Markle (1967) describes some of these subtleties involved in generating the needed information:

> Procedures for eliciting these data vary. Some testers prefer to talk to the student throughout the process, a procedure which, of course, renders the student's final performance suspect, if not invalid. Others prefer to query the student who hesitates or errs, leaving him to his own devices when no danger signals are apparent. The data which may be missed under this condition are exemplified by statements which some of us have heard often: "I know what you want here, but . . ." and "I see your point, but it seems to me . . . ." There are at present no firm rules. Each programmer has his own (p. 122-123).

The theoretic rationale for tutorial procedures appeared to be based on the assumption that observation of a single student is the best way to identify and remediate deficient instruction. Proponents claimed that observation of more than one student will overload the observer so important subtleties are missed (S. Markle, 1967). Furthermore, it was asserted that large group testing often suppresses individual candid reactions or the "stupid" question which underlies a major program problem; whereas, the more intimate tutorial situation will be able to elicit a greater amount of relevant feedback.

On the other hand, S. Markle (1967) and Paulson (1969) pointed out several limitations to the tutorial method: (1) it is costly and time consuming, (2) the subtlety and variety of techniques involved in

the tutor's or tryout editor's task make it difficult to describe and perform, (3) extreme vulnerability to atypical students, and (4) spurious inflation of learning and/or motivational. This last limitation means that the mere presence of the tutor might well be a reinforcing stimulus which spuriously inflates the student's motivational state and that any "tutoring" done by the tutor confounds the measurement of en route and criterion student performance.

With respect to development of a model for the present study, the literature on tutorial tryout and revision methodology has provided some valuable guidance. Nevertheless, two other strategies have been used which warrant analysis before the MK I model can be presented. Therefore, the next section of this chapter reviews the literature relevant to a second methodology of tryout and revision: the large group approach.

The Large Group Approach

Paulson (1969) defined the large group approach as tryout of a prototype instructional system with groups of twenty or more students with provision for recording and analyzing specific types of data. Paulson's summary of the advantages of this approach are paraphrased as follows:

1. It is often just as easy to obtain intact classes for tryout of prototype instructional units as it is to get individuals.

2. The instruction per se is more similar to the conditions of actual use than the tutorial approach. If instruction is relevant to a given class, the tryout may be embedded into the larger ongoing instructional system.

3. Students are not harassed by the necessity of commenting on their progress or learning difficulties, nor is their attention focused on the "trouble shooting" nature of their participation or the tentative, developmental nature of the instructional system being evaluated.

4. The data obtained via large group procedures are far less vulnerable to unique or idiosyncratic personal characteristics since the data are normally summed across students.

5. The larger data base increases the probability that correct decisions will be made on system deficiencies, e.g., which deficiencies warrant revision.

Essentially the logic underlying the large group approach is that the greater amount of data produced by this method provides more believable evidence on which to base costly and time consuming revisions.

Since the instructional system of interest in this study is a SLATE, it must be noted that the inherent complexity of the audio, visual, and other stimuli considerably increase the cost (in time and dollars) of developing revisions over that of a simple programmed text. With SLATEs, therefore, it may be difficult to justify a costly revision on the basis of feedback from a single student as is often done in programmed text development.

Since the purpose of formative evaluation is systematic remediation of deficiencies and since the methodology should generate information on what in the prototype is deficient enough to warrant revision, it would appear that a large data base is desirable for SLATE formative evaluation.

Interestingly, the large group with its correspondingly large data base has long been recognized as essential for summative evaluation

activities (Scriven, 1967). However, for formative evaluation the exclusive use of large groups has thus far not gained any appreciable acceptance.

Research by Vandermeer.--Vandermeer (1964, 1965) conducted two studies using large groups (intact classes) of school children to provide information leading to the improvement of a film (1965) and a filmstrip (1964). The methodology in both studies involved development of multiple choice instruments covering every informational aspect of the film or filmstrip. Following showing of the prototype versions to intact classes, test items showing the greatest difficulty (lowest student recall) were correlated with the specific part of the presentation where the information should have been learned. The researchers then revised the deficient portion to: (1) afford more cues or higher visibility (arrows, etc.) or (2) include less complex language in the narration. The revised versions were then shown to equivalent intact groups in other schools. The results were equivocal in both film and filmstrip studies. Some of the revisions "worked" and others did not, so the net effect was NSD. After revising a second time, Vandermeer showed a significant ($P < .05$) improvement on one-third of the items which reflected revisions in the film and one-half the items which reflected revisions in the filmstrip. The author did not discuss or advance reasons for the equivocation and nature of his results.

Of interest to the present study, however, was the fact that at no time did Vandermeer or his associates interview or personally observe any of the 203 Ss viewing the film or 216 Ss viewing the filmstrip. The sole basis for revision was test item data and experimenter post hoc analysis of the instructional stimuli. In light of Vandermeer's failure

to successfully revise, it would appear that to increase the probability of success, students must provide first hand feedback possibly including input on redesign of the deficient sections. Stated another way, test item errors may locate the troublesome part of a prototype lesson, but as shown in Vandermeer's research, "expert" post hoc analysis does not necessarily remediate student learning problems.

Discussion of the Large Group Approach

The large group tryout approach offers the advantages of generating large amounts of data to identify problems, but lacks the direct, specific input from students needed to develop appropriate revisions. With respect to development of a model of formative evaluation, the tutorial and large group techniques each have advantages and both are reflected in the model developed for this study.

An Approach Combining Individual and Group Data

By far, the most widely accepted approach to formative evaluation literature as reported in the literature is one which combines, in iterative fashion, data from both individual students and large groups. The paradigm for this approach is clearly illustrated by the flowchart in Figure 2 taken from Programed Learning: A Practicum, by Brethower, et al. (1966).

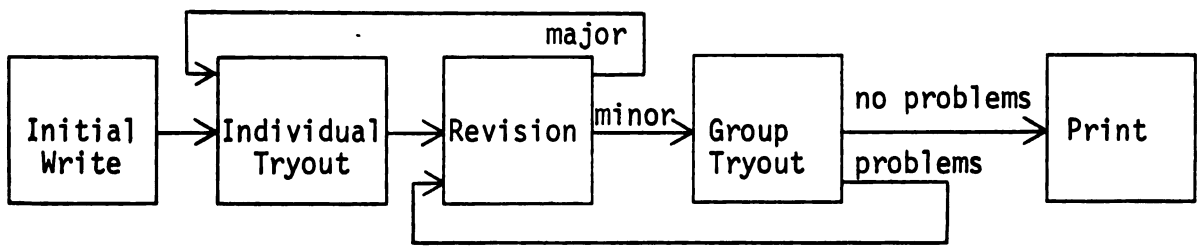


Figure 2.--Schematic Representation of the Recommended Testing-Revision Procedure (Brethower, et a., 1966, p. 169)

This approach is advocated by numerous other authors: Taber, Glaser and Schaefer (1965, p. 144-145); Pipe (1966, p. 56-59); Paulson (1969, p. IV-46-47); Schutz (1967, p. 21); S. Markle (1967, p. 111); and Briggs (1970, p. 172-173).

In addition, empirical studies have been done by D. Markle (1967), Anderson (1967), and Short (1968) each using this combined approach for tryout and revision procedures. Each of these studies resulted in statistically significant differences favoring revised over prototype versions of programs. For example, David Markle (1967) conducted a developmental study in which individual and group feedback was used not only to revise print and film materials, but to establish the course objectives, to determine the learning sequence, and to develop the evaluation instruments as well. In short, student feedback was used to support as many course development decisions as possible.

In Markle's study, the major objective was to develop a basic first aid course which, in seven and one-half hours, would exceed the performance of an existing ten-hour course. A set of test questions, derived from analysis of thousands of accidents, was defined as the course objectives and pre-tested on trained and untrained members of the student population. After removing the items known to nearly all the typical trainees, the remaining items became the first draft of the course. First, individual tutoring enabled Markle's development group to add instructional material gradually, as required, until students were achieving at the criterion level. The major basis for revision was error rate, response time, and prompting by the tutor. A similar procedure was used to develop a series of films starting with black and

white "still" pictures as a first draft film. Additional pictures, camera angles, color, and motion pictures were added as required, based on student feedback. After three to five students achieved criterion performance with little or no tutoring, the instructional sequence was then tried out with large groups (N=22 to 30) and revised until the group was achieving 90% of the criterion test.

The results of this study are truly exceptional and are summarized by Markle:

> These instructional engineering methods have resulted in the attainment of the proper objectives. In addition to the desired increase in efficiency as a function of decreased time, the new 7½ hour course is far more effective than the 10-hour standard courses with which it has been compared. On one wide-range test used for comparisons, untrained subjects achieved a mean score of 85, subjects trained in standard first aid courses achieved a mean score of 145, while subjects trained in the new course achieved a mean score of 270, out of a possible maximum of 326 points. Similar results were obtained with other tests and other subjects (p. 1).

Three inferences relevant to the present study may be drawn from Markle's work. First, instructional systems of greater complexity than a programmed text may be markedly improved by tryout and revision based first on tutorial tryouts and then large group tryouts. Second, this combined iterative approach seems very profitable when working with volunteer, adult students. Third, in agreement with the work of Mager (1961), students can provide a very significant input into the fundamental design of the instructional system when they are given the chance. They should be consulted as early as possible in the development process.

Discussion of the Combined Approach

Thus far, several research studies have been reviewed which seem to indicate that significant differences between original and revised versions occur most often when achievement data are combined with first

hand direct feedback from students. A widely used approach which combines achievement data and direct feedback involves tutoring a single student as he encounters problems and incorporating the tutorial instruction into the lesson. Although, the tutorial approach is the most sensitive to individual learning problems, this same sensitivity makes it highly vulnerable to atypical students and succeptable to embarking on costly and possibly non-functional revisions. Moreover, the task of the "tutor-editor" is difficult to perform and variability in tutorial techniques will affect the quality and quantity of data collected.

The large group approach provides a broader more credible data base and hence reduces the possibility of idiosyncratic revisions which might result from the tutorial method. Furthermore, the large group approach provides far more accurate measures of student learning as a result of the "program" or instructional materials. On the other hand, direct interaction with students is inhibited, and serious deficiencies may not be identified. In addition, a large amount of data can be generated which requires careful organization and display before it is usable.

The third approach, combining tutorial and group data in iterative cycles, appears to provide the best of both techniques and was adopted as the point of departure for development of the flowchart model in this study.

Related Methodological Issues

Although the overall approach to formative evaluation design has been selected, a number of related methodological issues remain. For example, the types of data collected in most of the studies reviewed

earlier were: (1) student achievement data and (2) observational (process) data. Are these two types of data sufficient for the present study, or should others be included in the model? If other types of data are needed, what specific indicators should be used?

Of direct relevance to these questions was the comprehensive treatment of measurement by Schalock (1969) in which he reviewed the strengths and weaknesses of various measures and the paper on evaluation of instructional systems by Paulson (1969) in which he analyzed the problems, needs, and alternatives available for formative evaluation. Of particular interest was Paulson's summary in which he suggested that certain specific measures were appropriate for providing given types of data. The relationships suggested by Paulson are paraphrased in Table 2.

Table 2.--Classes of Data and Specific Indicators for Formative Evaluation

Class of Data	Specific Indicators
1. ANTECEDENT DATA (assessment of student entry capabilities)	Pre-tests General abilities (standardized tests
2. TECHNICAL DATA (assessment of instructional stimuli quality)	Student comments Technical consultant comments
3. PROCESS DATA (assessment of students' behavior during the learning experience)	Tryout monitor observations and comments
4. LEARNING DATA (assessment of student progress towards learning objective)	En route responses and feedback during the lesson
5. CRITERION ACHIEVEMENT DATA	Post-test, criterion referenced
6. ATTITUDINAL DATA	Rating scale Questionnaire Student comments

The six classes of data shown in Table 2 seem to represent virtually every important aspect of a prototype lesson. Therefore, these six data types were included in the preliminary model of formative evaluation.

One final question which must be resolved relates to measurement of student achievement. The question at issue is whether student achievement should be measured against a specific standard (criterion referenced) or against performance of other students (norm referenced). The fundamental issue is: what type of information regarding student learning would be most useful for systematic remediation of learning deficiencies in prototype SLATEs?

Glaser's (1963) paper in The American Psychologist stimulated considerable interest in the kind of measurement that was suitable for assessing the quality of instructional enterprises rather than discriminating among individuals: He states:

> The scores obtained from an achievement test provide primarily two kinds of information. One is the degree to which the student has attained criterion performance, for example, whether he can satisfactorily prepare an experimental report, or solve certain kinds of word problems in arithmetic. The second type of information that an achievement test score provides is the relative ordering of individuals with respect to their test performance, for example, whether Student A can solve his problems more quickly than Student B (p. 374).

> Measures which assess student achievement in terms of a criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others (p. 375).

After showing that criterion levels can be established at any point from zero proficiency to perfection and assessed at any time during instruction, Glaser cogently argues that such criterion referenced achievement tests are far more useful in developing effective instructional treatments than tests which differentiate among individual students.

The logic underlying this argument is that prerequisite to improvement of instructional treatments is identification of what is wrong in terms of substandard student performance. Such substandard performance is most easily recognized by comparison against a criterion. Student achievement below the standard is, by definition, a deficiency in the prototype. Thus, to be maximally useful for formative evaluation, measures of learning must be defined in terms of observable pupil performance at or above a specific standard. For these reasons, the principle of criterion referenced measurement is reflected in the preliminary model of formative evaluation in this study.

Matrix Summary of the Literature Reviewed

A summary of the information obtained from the foregoing review of the literature is presented in Table 3. The majority of the methodological questions have been answered in a preliminary manner, so a primitive flowchart model may now be stated.

Formulation of the MK I Model

Formative evaluation requires several types of data which must be collected under three different conditions. It was recommended by Paulson (1969) that data on technical quality of the instructional stimuli and students' entering abilities be collected prior to instruction. The process, en route learning, criterion learning, and attitudinal data should be collected in both tutorial and large group situations. This gives rise to a three stage model shown in Figure 3.

Table 3. Matrix Summary of the Review of the Literature.

ISSUE	SPECIFIC QUESTIONS	AUTHORS REVIEWED: ROBECK	SILBERMAN & COULSON	HORN	DICK	VANDERMEER	MARKLE	BRETHOWER	PAULSON
Problem Identification	Data Types	Achievement & Direct Observa.	Achievement & Direct Observa.	Achievement & Direct Observa.	Tchr & Stu Comments Error Rate, Test Items & Text	Recall Test	Achievement & Direct Observa.	Achievement & Direct Observa.	See Table 2
	Instrument Types	Achievement Test Mult-Choice Items	Achievement Test Mult-Choice Items	Achievement Test Mult-Choice Items	Achievement Test Mult-Choice Items	Mult-Choice Items	Criterion Referenced	Mult-Choice Criterion Referenced	
	Sampling Procedures	Unknown-Used Single Volunteers	Unknown-Used Single Volunteers	Unknown-Recommends Volunteers	Intact Classes Unknown for Selection of Revision Designers	Intact Classes	Volunteers, Randomly Selected	Volunteers Randomly Selected	Random Selected Volunteers
	Administrative Procedures	Tutorial Procedures Unknown	Tutor Interrupted at Own Discretion to Provide Assistance	Emphasize Role of Student & Reduce Anxiety. Tutor after 3-4 Frames of Difficulty (1)	Posttest Given at End of Each Chapter in T Text	Gave Test Immediately After Viewing	Tutor Interpreted at Own Discretion Group Used Immediate Posttest	Tutorial Similar to Horn - Group Procedures N = 20	Tutorial Similar to Horn - Group Procedures N = 20-30
Problem Analysis & Revision Hypothesis Development	Data Reduction & Interpretation	Data Analyzed Intuitively	Data Analyzed Intuitively & Hypotheses Generated on the Spot	Data Analyzed & Hypotheses Generated on the Spot	All Data Supposed to Use 7 Decision Rules (3)	Identified Missed Test Items & Where Concept was Taught	Item Analysis Error Rate & Tutor Judgment	Item Analysis, Frame Analysis Item/Student Matrix	Item Analysis Item/ Student Matrix
	Assignment of Priorities	Unknown	Multiple Cases of Same Problem Resulted in Problem Revision	All Problems of Equal Priority		Number of Students Having the Problem	Number of Students Having the Problem	Unknown	Not Clear
	Revision Hypotheses	Based on Tutor Judgment & Feedback From Student	Based on Tutor Judgment & Verbal Interaction with Student	Revise on the Spot & Tutor Until Problem Remediated		Unknown	Based on Tutor Judgment & Feedback from Students	Based on Tutor Judgment & Feedback From Students	Based on Feedback From Students
Revision Design & Integration & Evaluation	Revision Design & Integration	Unknown	Included Tutorial Information in Revised Programs	Included Tutorial Information in Revised Programs	Unknown	Revised Film Segments Based on Post Hoc Analysis	Included Tutorial Information in Revised Versions	Include Tutorial Information into Revisions	Include Tutorial Information Into Revisions
	Evaluation of Revisions	Statistical Significance Between Original & 2 Revised Versions	Statistical Significance Between Original & Revised Versions	Unknown	Not Evaluated	No Significant Difference Between Original & Revised Versions	Statistical Significance Between Original & Revised Versions	Unknown	Recommends Experimental Comparison

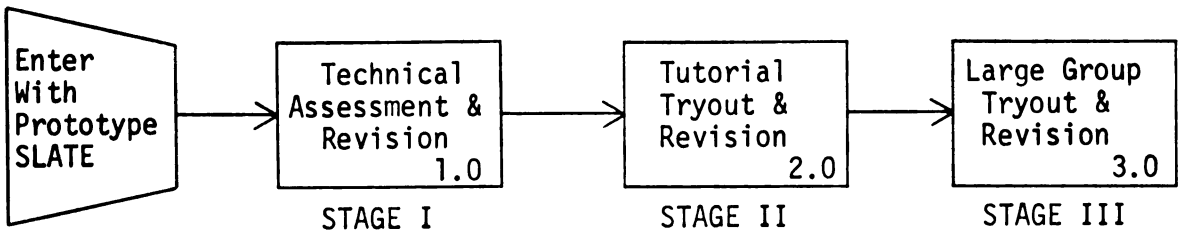


Figure 3.--Major Stages in MK I Model of Formative Evaluation

Stage I.--In recognition of the need to predetermine the technical accuracy of subject matter content, the technical quality of presentation media, and adequacy of evaluation instruments--before students interact with the instructional stimuli technical assessment is conducted.

Stage II.--Tutorial tryouts are intended to provide data on specific learning and communication problems which are best gathered during more intimate tutorial sessions than with larger groups.

Stage III.--In recognition of the limitations of tutorial tryouts, e.g., vulnerability to atypical Ss and author reluctance to base expensive revisions on data from a single student, large group (N=20) tryouts are conducted to broaden the data base.

A flowchart showing the MK I model at the first level of detail is shown in Figure 4.

A flowchart of the MK I model showing the fourth level of detail is depicted in Figure 5. This flowchart represents the final configuration of the MK I model developed from this review of the literature. In the next chapter, assessment and revision of the MK I model are described.

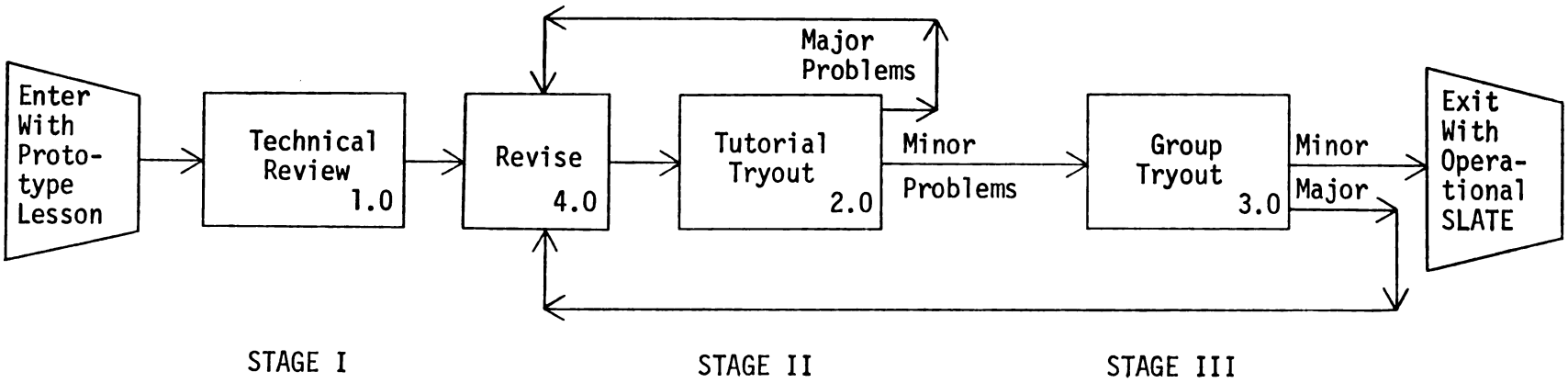


Figure 4.--Configuration of the MK I Model of Formative Evaluation Showing the First Level of Detail

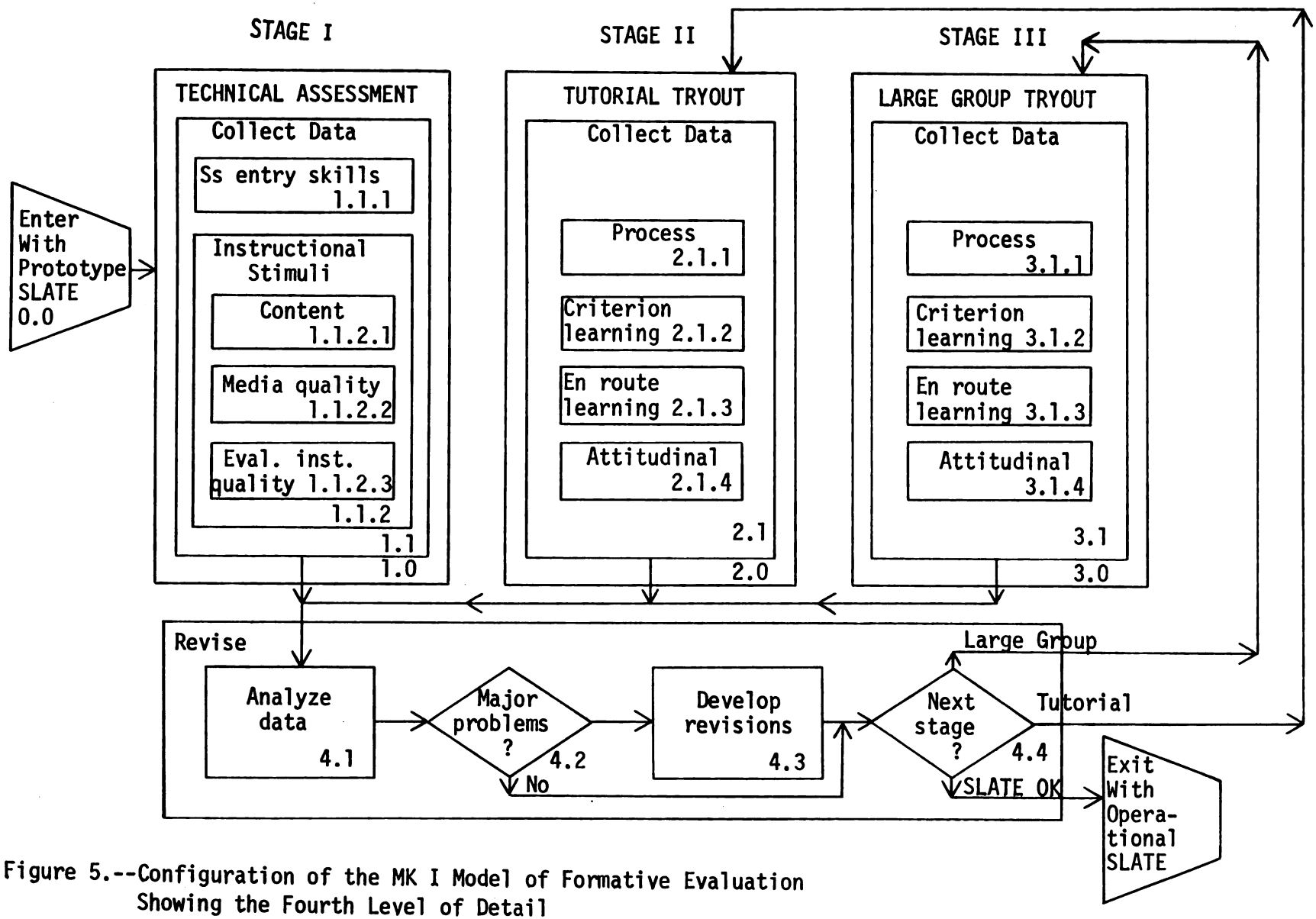


Figure 5.--Configuration of the MK I Model of Formative Evaluation Showing the Fourth Level of Detail

CHAPTER III

ASSESSMENT AND REVISION OF THE MK I MODEL

Introduction

The MK I model, having been developed from a review of the literature, is essentially an idealistic statement of "what ought to be" rather than what is practicable. The model, in the broadest sense, is an untested hypothesis. Therefore, a major objective of this study is validation: to obtain feedback on the model in settings as close to the actual using situation as possible so that improvements can be made.

Validation was carried out in two phases. Phase one, which is described in this chapter, involved interviewing seven selected SLATE authors to determine their opinions on the practicability of the model and its congruence (or lack of it) with their personal tryout and revision procedures. On the basis of these interviews the MK I model was revised to take account of realistic constraints and reflect a compromise between theoretic and pragmatic considerations.

The second phase of validation involved using the revised (MK II) model in a series of field trials; the goals of which were to: (1) revise prototype SLATEs, and (2) further refine the MK II model. The methodology and results of these field tests are reported in chapters IV and V respectively.

Overview

This chapter describes the first assessment and revision of the MK I model. Structured interviews were conducted with seven SLATE authors in which a comparison was made between MK I procedures and those used personally by the authors. Interview questions were developed by extrapolation from the issues presented in Chapter II (Review of the Literature). Respondents were six college faculty and one R & D specialist selected on the basis of their previous experience with SLATE development. The experimenter personally conducted all interviews and summarized these data by means of a narrative analysis of responses to each question. This analysis revealed several major problems with the MK I model which lead to the development of MK II "mini" and "maxi" models. Both MK II models prescribed identical steps, which differed only in level of detail. The "mini" version was developed in response to a need for a simplified model to introduce the total process to less experienced faculty. The "maxi" was used only by the experimenter.

The remainder of this chapter is organized as follows: (1) a description of interview procedures including selection of respondents and questionnaire development; (2) a narrative summary of responses to each question; (3) discussion, conclusions, and development of changes to the MK I model; and (4) description of the MK II "mini" model and flowchart of the MK II "maxi" model. (A description of MK II "maxi" model is in Appendix B.)

Procedures

Questionnaire Development

An interview questionnaire (Appendix A) was developed by the experimenter (E) through extrapolating questions from each of the issues

addressed in the review of the literature. Basically, the questions were organized around six factors. The factors and specific items related to the factors are shown in Table 4.

Selection of Respondents

Since the purpose of the interview was to collect data from experienced practitioners, respondents were carefully selected using the following criteria:

1. Each respondent was personally acquainted with the experimenter so a minimal level of trust and confidence had previously been established.

2. Each respondent had been the major author of several multimedia self-instructional lessons which had been used at least one term/semester within the previous school year.

3. Each respondent expressed a willingness to discuss their tryout and revision procedures with the experimenter, either in person or telephonically.

4. The sample would be stratified in terms of respondents' grade level, subject matter, affiliation, previous experience in teaching and SLATE design, and overall "sophistication" with regard to instructional technology and development. Operationally, sophistication was defined to mean the distinction between professional program developers--those respondents who derive their living from developing instructional materials--and non-professional program developers--those respondents whose main responsibility is that of a teacher/faculty member and whose developmental efforts are normally in support of their own specific "course."

A list of fifteen potential respondents was drawn up from which seven were selected. These seven were selected mainly on respondents' willingness to be interviewed on this topic and their proximity to MSU. Table 5 summarizes background data on each respondent in the sample.

Interview Procedures

Six respondents were interviewed by the experimenter (E) in their office. In one case, due to geographic distance involved, the interview

was conducted via telephone. A tape recorder was not used as two of the respondents indicated some concern regarding the recording of their comments.

Table 4.--Factors Used in Questionnaire Development

Factor	Items
1. Determination of data types, sources, and units of analysis appropriate for formative evaluation of SLATEs.	4, 5
2. Determination of appropriate methodologies for collection and analysis of required data.	3, 2, 6, 8, 9, 14
3. Development of systematic procedure for assigning priorities among instructional problems and alternative solutions.	7, 10, 11
4. Development of strategy for systematic design of revisions.	12, 13
5. Practicability of MK I model.	17 a, b, c
6. General assessment of the formative evaluation process.	16, 15

Table 5.--Background Data From Respondents

Respondents	Affiliation and Position	Degree & Teaching Experience	Previous Training or Experience in SLATE Design	Instructional Systems Using the SLATEs
R_1	MSU Professor	Ph.D. 17 Years Teaching	No previous exp. Had some prof. help during SLATE devel.	3 cr. course in An. Hus. Lecture + lab + 14 SLATEs. 175 freshmen
R_2	MSU Instructor	M.A. 4 Years Teaching	1 yr. train. with Media Inst. Had some prof. help during SLATE devel.	5 cr. Biochem. course Lecture + lab + 27 SLATEs, 200 sophomores
R_3	MSU Professor	Ph.D. 15 Years Teaching	None--but had moderate prof. help during SLATE devel.	3 cr. course in Soil Science. Lecture + lab + 11 SLATEs, 50 sophomores
R_4	MSU Professor	D.O. 6 Years Teaching	None--but had extensive help during SLATE devel.	3 cr. course in Vet. Med. 10 SLATEs in prep. lab., 20-30 stu.
R_5	Dept. Chairman Lansing Comm. College	Ph.D. 22 Years Teaching	None--little prof. help during devel. of SLATEs	Two 3 cr. courses, 9 SLATEs ea. in Phy. Sci. & Geol., 250 freshmen
R_6	MSU Assistant Professor	Ph.D. 5 Years Teaching	1 yr. as grad. teaching assistant with Postlethwait Had some help during SLATE devel.	5 cr. Bio. Science course. 14 SLATEs 500 sophomores
R_7	USOE Regional Lab. Staff member	Ph.D. 3 Years	Wrote several PI texts Worked with several R & D projects	K-5 elemen. Lang. Arts program. 124,000 students/sem. 36 modules

Interview Data

The data collected during the seven interviews are presented by summarizing responses to each question. This method should provide readers a better insight into the problem of formative evaluation and the rationale underlying changes in the MK I model.

Discussion of Individual Questions

Question 1: DESCRIBE THE SUBJECT MATTER, TARGET POPULATION, AND INSTRUCTIONAL SYSTEM IN WHICH YOUR SLATES WERE USED.

The responses to this question, as well as other background data, are summarized in Table 5.

Question 2: DID YOU REVISE YOUR MATERIALS AFTER A "FIRST DRAFT" OR PROTOTYPE HAD BEEN PRODUCED? BEFORE OR AFTER FULL SCALE USE WITH THE TARGET POPULATION?

Six respondents indicated they had revised their SLATEs beyond the "first draft" stage. The degree of revision varied from major to minor; the majority reported only minor revisions. The definitions of "minor revision" varied considerably in terms of specific activities but usually involved less than five man-hours of author effort.

One respondent (R_4) indicated he had not revised at all--primarily because the original versions seemed adequate to achieve the intended SLATE objectives (i.e., to reduce laboratory time and decrease frequency of fatalities during surgical procedures with animals).

Of the six respondents who revised their units, only one respondent (R_7) had conducted formative evaluation in the sense of having revised SLATEs before extensive student utilization. The others had revised only after extensive use by students (one term or longer) had revealed serious problems in the prototype versions. The one author who did revise was an

instructional research and development specialist who did so because of "company policy" and the fact that the specific project involved a large potential loss of prestige and federal funds if it failed.

Question 3: WHAT WAS YOUR REVISION STRATEGY? DID YOU HAVE A PREDETERMINED PLAN? IF SO WHAT WAS IT AND HAD YOU USED IT PREVIOUSLY? IF YOU HAD NO PLAN, HOW DID THE REVISIONS EVOLVE?

Of the seven respondents, only the professional had a predetermined strategy for revision development. Although the professional was well aware of specific techniques and desirability of formative evaluation, he cited a frustrating inability to apply these techniques in many cases due to the enormous commitment of resources required. The economic constraints were cited as the main factor precluding his commitment to the formative evaluation process. On the other hand, there seemed to be a sliding scale of project importance, wherein a given project if it involved sufficient prestige and dollar cost, automatically warranted formative evaluation.

The professional felt that the entering capability and experience as a program/SLATE author was a critical variable in determining the need for formative evaluation. That is, the greater the experience of the author, the less the need for formative evaluation--hence this procedure was probably most relevant to novice or non-professional SLATE authors. Paradoxically, the non-professional program authors in this sample were not aware of a need for any type of tryout-and-revise developmental strategy.

Of the six non-professional SLATE authors, not one had a systematic plan for revision, nor were they aware of the desirability of

revising SLATE materials before large scale student use. The most frequent rationale for this non-interest in SLATE tryout and revision was as follows: (paraphrased) "I don't revise my lectures, my labs, or my textbooks before use with my class, so why should I spend valuable time revising my SLATEs?" All six admitted the likelihood that prototype SLATEs might be deficient in some important respects, but since SLATEs were a subsystem of the "class" they personally were teaching, they felt that intact class usage was a justifiable method of prototype tryout (e.g., they could correct SLATE deficiencies during lectures).

Question 4: FROM WHOM DID YOU OBTAIN FEEDBACK: INDIVIDUAL STUDENTS, GROUPS, EXPERTS, OR OTHERS?

The respondents who did revise SLATEs obtained feedback both directly and indirectly from students in the target population. Direct feedback was obtained most frequently from individual students complaining personally to the respondent either after lectures or while the respondent circulated through the carrel room while the student was using the SLATE. Indirect feedback was obtained from lab assistants, carrel room monitors, or discussion group leaders who reported serious discrepancies sometime after they occurred. For example: (paraphrased) "SLATE on X is really bad--the students are not finishing in one hour; they cannot do the workbook problems; the experiment consistently fails--etc." No systematic sampling or student selection procedures were used except in the case of the "professional" working on the large federal project mentioned earlier.

In two cases, scripts were read by colleagues for content accuracy, but the majority of feedback was obtained randomly from students via intermediaries such as graduate teaching assistants (GTAs) in the course.

Question 5: WHAT KINDS OF FEEDBACK DID YOU TRY TO GET?

Six respondents actively tried to gather student attitudinal data; while all seven respondents made efforts to assess student achievement (learning) from the SLATEs. In two cases (R_2 and R_7) the student background/demographic data were collected. In most cases, observational data were randomly collected via respondents' personal visits or through intermediaries such as GTAs. In sum, none of the respondents used more than two types of data.

Question 6: WHAT METHODOLOGY WAS USED TO GATHER THE VARIOUS TYPES OF FEEDBACK?

In all cases, respondent-designed measures were used to assess formally student end-of-SLATE learning and attitudes. In six cases, learning measures were typical paper-and-pencil tests using true-false and multiple choice items. In two cases, however, performance in post-SLATE "action" laboratories was the prime source of data on student learning.

Attitudinal measures were used by six respondents but collected at different times. For example, three respondents collected attitudinal data following each individual lesson. The others collected attitudinal data at the end of the course.

Process data were collected the least systematically. In most cases the methodology involved random observation/interaction between the respondent (SLATE author) and his students as he visited the carrel room during operation. The major source of process data was verbal report from intermediaries such as carrel room attendants or lab assistants who described student difficulties with specific SLATEs to the respondent--after a number of students had been observed having a similar difficulty.

The number of similar discrepancies which constituted a reportable incident varied considerably, but respondents agreed a reasonable estimate would be more than 20% of the students using the SLATE.

Question 7: HOW WOULD YOU CLASSIFY THE TYPES OF PROBLEMS YOU FOUND? ADMINISTRATIVE/TECHNICAL? COMMUNICATION? LEARNING OR TASK RELATED? (EXPLAIN THESE CATEGORIES.)

This was not a particularly useful question as most respondents did not understand the categorization system and often a lot of time was wasted in explanation. What it did show was the complexity of the MK I model's classification scheme. Nevertheless, a number of respondents stated that many problems fell into the administrative/technical class such as waiting in line due to limited space; equipment malfunctions; not reading or following directions; necessary equipment becomes lost, misplaced, and mislabeled. Communication and message design problems appeared next in frequency, with boredom and general inattention cited most often as the primary problem. Learning and task related variables were hardly mentioned unless prompted by the interviewer. Non-professional respondents seemed to assume that variables such as the learning objectives, evaluation instruments, the sequence, organization, response type and frequency (if any) were "given" and not subject to modification. In about one-half of the cases, respondents had given serious consideration to the student's response and feedback, both the type and frequency, while in the other half little consideration was given, and SLATEs were simply illustrated lectures with a post-test.

Question 8: HOW DID YOU SUMMARIZE, DISPLAY, AND ANALYZE YOUR DATA?

Two types of data were formally analyzed: student achievement and attitudes. Achievement data analysis was done primarily by frequency

counts of missed items and conventional item analysis. Attitudinal data were normally summarized as a percentage response on individual items. Consultants were used only in two cases to analyze the data. Most frequently, the designers seemed to make a subjective decision as to whether a problem warranted revision.

Question 9: HOW DID YOU DETERMINE IF A REVISION WAS REALLY NECESSARY? HOW MUCH OF WHAT TYPE OF DATA WERE NECESSARY FOR COMMITMENT OF RESOURCES (TIME AND DOLLARS) FOR REVISION?

No one particular data source or criterion seemed to emerge. Each respondent seemed to have an intrinsic weighting system which included personal observations, attitude and examination error data, and personal time and dollars available at the time. It did not seem as if any one data soruce was sufficient, but rather that multiple sources would have to correlate before revision action would be warranted. However, any action depended on how costly the revision seemed to be.

For example, a bad examination item (missed by a lot of people) could be revised quite easily by cutting a new stencil--unless the item was embedded in a large workbook. If the latter were the case, the item remained unchanged and an erratum sheet was posted in the carrel room; i.e., "disregard item X; it is no good."

It was obvious that audio visual materials were revised very reluctantly. The time and costs were considerable and designers simply did not have time or dollars to revise any but the most deleterious materials--and then only after one or two terms had gone by and the data was overwhelming. Interestingly, when a faculty member was teaching the same course in which the SLATE was used, he could compensate for failures of

one SLATE by reteaching in lectures or quiz sections the content which students did not learn from the deficient SLATE. Also, laboratory assistants served a tutorial function to reduce the seriousness of learning difficulties that were the result of ineffective prototype SLATEs.

Usually, the criterion for revision was multiple inputs which indicated that a specific problem existed in a SLATE; that is, if the lab assistants consistently reported a problem, and/or if the designer personally observed the problem; and if student achievement data reflected a deficiency--then the decision was made to revise the SLATE.

Question 10: WHICH COMPONENT DID YOU REVISE: OBJECTIVES, EVALUATION INSTRUMENTS, AV MATERIALS, OR SOME OTHER?

In the six cases in which revisions were made, a combination of all three components (objectives, evaluation instruments and AV materials) were revised. This was due to the highly interdependent nature of the instructional stimuli in SLATEs; e.g., revision of one component necessitated revision of other components.

All seven respondents indicated they would have liked to revise the SLATE objectives, but only three did so because of the magnitude of this undertaking. Of the three who revised objectives, only one added objectives whereas the other two deleted objectives.

Other aspects of the SLATEs which were revised were sequence and complexity of lab experiments. In one case, the overall sequence of SLATEs was revised to allow for better integration with lectures and laboratories. In another case, a complex lab experiment which failed 50% of the time was replaced with a more simple and reliable one.

Question 11: HOW DID YOU DETERMINE THE DESIGN OF REVISIONS?

None of the respondents, including the professional, had a systematic approach or theoretic position on AV materials design. Invariably, they used an intuitive approach to revising AV materials similar to the approach used in the original design process.

Revision of evaluation instruments consisted largely of deleting items of low difficulty or discrimination or rewriting item stems or foils to reduce ambiguity.

Revision of objectives was largely based on an intuitive decision regarding course content and/or difficulty level of the course concepts. It appeared that the respondents typically overestimated the students' entering capability, and consequently, revised objectives were simplified versions of the original ones.

Question 12: HOW MANY REVISIONS (CYCLES) DID YOU MAKE?

The responses to this question varied considerably as most respondents made some revisions to most, but not all, of their SLATEs. For example, one respondent did not revise at all. Of the six who did revise, most made one set of minor revisions (less than five man-hours or author work) on each SLATE. However, several respondents indicated they had, after several cycles of "patch up" minor revisions, made major revisions amounting to a complete redesign and reproduction of a given SLATE.

Each SLATE/author combination was a unique case, and the number of revisions conducted appeared to be a function of how good (or bad) the prototype was, the resources available, and the institutional press on the individual author. It did not appear, however, that the

respondents were willing to commit large amounts of time and resources to a major revision until a large amount of data had been accumulated, over approximately one year's time.

The professional, on the other hand, indicated that four cycles of revisions were accomplished before the final SLATEs were widely distributed. The last two revisions were based on field tests where the authors were not present during student use and all data were from end-of-SLATE test scores and teacher interviews.

Nevertheless, to most respondents the technique of multiple iterative revisions did not seem feasible due to its high cost.

Question 13: WHAT PERCENTAGE OF PROTOTYPE DEVELOPMENT DID YOU SPEND ON TRYOUT AND REVISION?

The time range was from 0 to 200%, with the average about 20%-30%. When queried as to what was the original time investment on a pre-SLATE basis, most were unable to recall as various SLATEs were being developed simultaneously. Furthermore, the various SLATEs took different lengths of time to develop depending on the author's teaching load, complexity of the unit, previous preparation, whether materials were already used in class, the production capability of the author, and other situational variables.

After some probing, the experimenter (E) extrapolated an average development time of between 50 and 100 author hours per prototype SLATE--exclusive of support man-hours (typing, collating, photography, sound recording, etc.).

Question 14: WAS IT WORTH IT? HOW DID YOU DETERMINE IF THE SLATE(S) WERE IMPROVED?

All respondents who did revise were absolutely certain the revisions were effective but had little objective evidence on which to base these judgments. Not one of the non-professional respondents made statistical comparisons between achievement test scores or attitudinal scores on original and revised versions. This technique was felt to be too time consuming, and the relative effectiveness of a given SLATE could be determined through informal means such as lab assistants, GTAs', and students' questions in lectures.

On the other hand, the professional did use statistical tests to compare original and revised versions on measures of student achievement, student attitudes, and teachers' attitudes.

In general, the respondents did not seem concerned with an objective evaluation of their revisions. With them it was simply a foregone conclusion that the revised versions would be an improvement over the prototype.

Question 15: IF YOU HAD TO DO IT OVER AGAIN, WHAT WOULD YOU DO DIFFERENTLY?

Many replied that in retrospect they would select different objectives and/or content; e.g., their original objectives were overly optimistic in terms of students being able to achieve them in SLATEs. This may be a reflection of poor program design as well as curricular refinement. Several respondents commented that the idea of "revision as you go" sounded like a good one, but there seemed too little time to perform the tutorial procedures.

Another major problem in SLATE revision was selecting students with necessary entry skills. Most of the respondents' SLATEs were

embedded in a larger instructional system--a "course"--and were in large degree dependent upon students obtaining necessary prerequisites from earlier SLATEs as well as from other course related learning experiences. Naturally, the later a SLATE was to be used in the course sequence, the more serious was this problem of obtaining students who possessed the prerequisites yet were naive with respect to specific SLATE objectives.

This difficulty, along with the time and expense inherent in tryout and revision procedures, tended to reduce interest in formative evaluation as represented in the MK I model.

In sum, the majority of respondents indicated they would change the subject matter content of their SLATEs and attempt a closer integration between SLATEs, but the overall process of SLATE development would remain basically unchanged.

Question 16: DO YOU THINK THE MK I MODEL IS PRACTICABLE? IF NOT, WHY NOT? WHAT CHANGES WOULD YOU SUGGEST TO MAKE IT MORE PRACTICABLE?

Without exception, all respondents stated that the MK I model was highly impractical in the "real world" and they would be unwilling or unable to use it. Several reasons dominated. First, the model seemed overly complex and time consuming. (It appeared to E that the flowchart itself simply overwhelmed respondents.) Second, the concept of iterative revisions based on tutoring single students appeared totally out of the question from the standpoint of data credibility and cost effectiveness. In other words, given the extremely high development costs of SLATEs (both labor and materials) and the difficulty of integrating slides, tapes, workbooks, models, laboratory exercises, directions, etc., authors simply will not revise this whole logistical system on the basis of feedback from one student.

On the other hand, the prospect of revising on the basis of group feedback seemed more acceptable, but logistical and sequencing difficulties posed serious problems. That is, SLATEs in highly technical areas such as biochemistry, soil science, geography, and medicine, are highly interdependent and must be hierarchically sequenced. This means that student tryouts must follow the same hierarchical sequence which poses major logistical difficulties in terms of coordinating design, production, tryout sequencing, and class sequence. SLATE production must coordinate with learning activities within the "class" embedding the SLATEs so that students who have the necessary prerequisite knowledge can be obtained at the proper time in the course sequence. If a moratorium is declared so that the class is not offered while SLATEs are being developed, available students may not have prerequisites. Although SLATEs are supposed to be self-instructional, they nevertheless depend to some extent on lectures, text, and lab sessions of the embedding course. If there is no ongoing course from which students may be solicited, a suitable sample for formative evaluation may be impossible to obtain.

The major changes to the MK I model suggested by respondents were: (1) deletion of the tutorial tryout and revision phase, (2) deletion of the technical assessment and revision phase, and (3) development of a logistical/sequencing procedure which would allow group tryouts to be optimally sequenced within an ongoing course.

The procedures contained in the technical assessment and tutorial phases were recognized as potentially valuable but not worth the effort. For example, most respondents seemed very reluctant to allow peer review of their "rough draft" prototype work either for technical or stylistic

comments. Most regarded themselves as "content" experts; hence additional technical review was redundant. In addition, most felt they were capable of assessing media and evaluation instrument quality due to previous experience teaching.

With regard to deletion of the tutorial tryout and revision phase, most respondents felt that basing SLATE revisions on feedback from a series of individual students did not seem feasible or cost effective. SLATEs were too complex and costly to revise on the basis of one or two students. Furthermore, the tutorial procedure appeared excessively time consuming from the author's standpoint and excessively costly if revisions were to be made after each student's tryout.

On the other hand, most respondents seemed agreeable to use of a large group tryout procedure which would quickly generate a large amount of data.

Discussion of Interview Data

Several trends clearly emerged from these data. First, none of the respondents, except the professional, felt that revision prior to full scale use was warranted due to press of time and lack of resources. Second, tutorial tryouts with individual students did not seem to be a feasible technique or basis for revision.

The complex and highly coordinated nature of SLATE instructional stimuli (slides, slide change signals, audio tape content and directions, workbook, student responses, knowledge of results, etc.) made it very difficult to change anything once the prototype was set up so the first student could use it.

In recognition of this situation a heuristic clearly emerged; namely, SLATE authors need a rather overwhelming amount of data to convince them that any revision effort is "worth it." Operationally this means that several students must have encountered a given problem, and that more than one data source must have corroborated the same problem (such as personal observation and post-test errors) before revision action is taken. Furthermore, several revisions must be required on the same SLATE before any action is taken. In other words, the vehicle must have several serious discrepancies before it warrants an overhaul.

As represented by this sample of SLATE authors, a very clear pattern of revision activity emerged. Typically, the SLATE was designed as well as possible. Then it was used in prototype form by the intact class under control of the SLATE author. During this initial usage, random feedback was obtained via authors' personal observations, verbal reports from lab assistants, carrel room attendants, discussion group leaders, and/or students. Systematic feedback was obtained from end-of-course evaluation of student learning and attitudinal data, and in some cases, assessment of student achievement and attitudes after each SLATE. Typically, however, the instruments used to collect data were too general to provide specific guidance for the design of revisions. Nevertheless, data on problems in various SLATEs gradually accrued from several sources. When sufficient corroborative data was obtained, and if time and resources permitted, revisions were attempted. These revisions were developed on an intuitive basis, often in consultation with GTAs (What should we do about "X?") but seldom, if ever, using the students as a source of design information. The most common

revisions reported by respondents was a reduction and simplification of subject matter content--a reduction in "coverage"--which reduced the average student time in the SLATE by 10%-25%. This differed from findings in programmed instruction studies where revised programs are usually longer than original versions.

It appeared that the major impact on the SLATE author of typical after-the-fact feedback data was a rapprochement between estimated and actual entering student capabilities and a reassessment of objectives and content coverage in given SLATEs. Typically, prototypes were too ambitious; so when revisions were made, the net effect was to reduce their complexity. Thus, feedback most often caused reformulation of course/SLATE content and objectives as well as revision in programming and/or presentation techniques. The regrettable aspect was the large number of students who were subject to the prototype versions until the author recognized what should have been in the SLATE and took appropriate action.

Conclusions From the Interview Data

These data clearly showed that with faculty similar to those interviewed, the MK I model was not practicable. MK I did not correspond even remotely to current practice, and of the seven respondents interviewed, none was willing or able to use the model in its present form. A major reason given was that it was logistically impossible to coordinate design and production with selection and conduct of tutorial revisions followed by large group tryouts and revision. The major problems were: (1) obtaining naive students at the proper time, (2) author "release" time, and (3) revision costs. While most respondents conceded

that the use of the MK I model would likely result in better SLATEs than they currently had, none felt that SLATEs needed to be that good; e.g., there were other important aspects of the course, and SLATEs per se simply did not warrant all that effort.

While the MK I model was designed to reduce uncertainty regarding formative evaluation, it seemed to raise more questions than it answered. Moreover, MK I did not recognize the severe time and financial constraints which operate in the practitioner's world, nor did it recognize certain characteristics of SLATE authors which inhibit formative evaluation. For example, university faculty typically regard themselves as subject matter experts and highly proficient teachers; consequently, they do not recognize a need to tryout and revise SLATEs before using them on their intact classes. The respondents felt that teaching and designing SLATEs were complicated enough without introducing more complexity and uncertainty by evaluating their SLATEs and possibly getting a bad report. Furthermore, several respondents were very reluctant to allow students to criticize their SLATEs, particularly in a face-to-face tutorial situation.

These data led to the conclusion that the concept of formative evaluation itself (basing revisions on feedback from students) must be "accepted" before any model is practicable. Assuming acceptance of the concept, then three major revisions of the MK I model can be inferred from the data: (1) logistical and conceptual simplification, to include either deletion or modification of the technical assessment and tutorial phases; (2) some procedure for reducing, during formative evaluation, the interdepency of SLATE instructional stimuli which dissuade authors

from changing anything (e.g., components are so highly interrelated that the smallest change becomes a major task); and (3) attention be given to obtaining corroborative data on major instructional problems so authors will be more likely to take necessary remedial action.

In conclusion, the MK I model must be simplified; its fundamental concepts justified to SLATE authors; the hierarchical interdependence of instructional stimuli reduced; feedback techniques must generate corroborative data.

Revisions to the MK I Model

Simplification

The model would be greatly simplified if the first two phases were simply eliminated leaving only the group tryout and revision phase. Data from the interviews supported such a move. However, it is the experimenter's opinion that the technical assessment phase is not sufficiently complex or time consuming to warrant complete deletion. Furthermore, it had been the personal experience of the experimenter that prototype SLATE evaluation instruments often were either lacking altogether or of such low quality that the necessary types of data for formative evaluation could not be generated. Therefore, despite the interview data, modifications to the MK I model did not include a deletion of the technical assessment phase.

Obtaining Corroborative Data

Some provision must be made in the revised model to obtain a sufficient amount of corroborative data so authors know what must be revised.

Based on the unequivocal response of practitioners, tutorial procedures should be eliminated. But the group tryout as it was formulated was not likely to generate the detailed information needed for identification and remediation of critical learning problems. In other words, tutorial procedures identified and solved learning problems while large group procedures were normally limited to problem identification. Therefore, some technique must be found which generates both the tutorial and large group data types. The critical aspect of this new technique is that it must generate a large amount of relevant and corroborative data with minimal expenditure of designer or student's time. This procedure must also be logistically compatable with an ongoing course context.

Group Debriefing as a Feedback and Problem-Solving Technique

E devoted considerable effort to the development of a technique which combined the tutorial and large group data collection potential, yet did so in a minimal length of time. While searching for a solution to this problem, E was struck by the functional similarity between formative evaluation and military debriefings. For example, it is common practice in the military to evaluate mission and training procedure effectiveness by means of formal debriefings. Usually, operations and support personnel participating in an exercise or training program are interviewed immediately following each mission to determine specific successes and problem areas. Information is collected from all participants, summarized, and formally reviewed by mission/training directors to determine how to improve mission effectiveness. Thus, first hand

information from operational level participants is fed back to the planning and design personnel. The function of formative evaluation is very similar; information on specific success and problems of participants (students) is fed back to the lesson author for purposes of improving the lesson.

In light of this similarity, it was reasoned that if SLATE formative evaluation were conceptualized as a "one shot" small group debriefing following a lesson, that data collection would be simplified.

Furthermore, some data have shown that during training program development when trainees participate in a mission debriefing, they not only provide planners with information on problems, but can often provide solutions to such problems. For example, when the U.S.A.F. Air Defense Command radar intercept system was developed, personnel operating the system in a training status were debriefed after each major exercise. In this way, critical problems were identified and remediated (Alexander, et al., 1962).

Reconceptualizing the Problem

It was reasoned that if formative evaluation were reconceptualized as tryout and revision by means of a small group debriefing/problem solving process, not only might identification of major discrepancies be facilitated, but quite possibly the debriefing might suggest more effective solutions than would otherwise be possible.

Development of Group Debriefing/ Problem Solving Procedures

In the present study, formative evaluation was reconceptualized as a group debriefing/problem solving process. The major source of

feedback on instructional problems was to be a group of students who were given the dual task of problem identification and development of solutions to the problems identified. It became necessary, therefore, to develop procedures appropriate to achieve these objectives.

Review of Literature on Group Processes

Much of the current research on group processes appears to have grown out of two separate but related historical movements. One movement emerged from the works of John Dewey who emphasized the social aspects of learning and the role of the school in training students for problem solving and for democratic, rational living (Schmuck & Schmuck, 1971). The other movement emerged from the empirical research of Lewin and the subsequent development of researchers and practitioners in the field of group dynamics (Bany & Johnson, 1964). The latter movement emphasized the collection of empirical data which supported the philosophical work of Dewey and introduced specific procedures for improving group processes (Bradford, Benne & Gibb, 1964).

During the past twenty years there has been an extensive accumulation of scientific research on small groups as the study of group dynamics developed as a subdiscipline of social psychology (Schmuck & Schmuck, 1971). In 1955, for example, Hare and others annotated a bibliography of 584 items on small groups. By 1959 Raven published a handbook which included 1385 references, and in 1966 McGrath and Altman published a bibliography of 2699 references. In addition, shorter analyses of group dynamics were published showing both the interest and magnitude of research in this area (Golembiewski, 1962; Luft, 1963; Olmstead, 1959; Shepard, 1964).

One trend in education resulting from research on group dynamics was the direct application of group research for the improvement of personal learning and/or for learning organizational processes (Schmuck & Schmuck, 1971). For example, one notable application was the technique for educating adults referred to as the training group (T-group). This technique was developed by the National Training Laboratories: Institute for Behavioral Science. Important publications relevant to the T-group were Bradford, Gibb, and Bene (1964) and Schein and Bennis (1965). Refinements to T-group technology grew out of research on organizational group processes (Katz & Kahn, 1965; Likert, 1961; March & Simon, 1958).

Until recently, much of the research in group dynamics has been done in industry and government rather than in school contexts. Lately, however, there has been an increased emphasis on the application of group processes to educational settings. The 59th volume of the National Society for the Study of Education (Henry, 1960) provided a social psychological theory on classroom groups and proposed ways of using research findings to improve instruction. Several recent works review empirical data on group processes in the classroom and other school settings (Bany & Johnson, 1964; Glidewell, et al., 1966; Lippett, Fox & Schmuck, 1964) while other works utilize data on classroom group processes to make recommendations for improving teaching (Schmuck, Chesler, & Lippitt, 1966; Fox, Luszki, & Schmuck, 1966; Chesler & Fox, 1966; Amidon & Hunter, 1966).

Emerging from this large accumulation of research was the recognition that a number of complex variables dynamically interact in any small group. Since the present study was concerned primarily with formative evaluation, no attempt was made to formally investigate the

numerous variables known to operate in group dynamics. Instead, the MK II group debriefing/problem solving procedures were based on generalizations drawn from previous research on group processes. Three works were the primary references for development of the MK II group debriefing procedures: Maier (1963), McGrath and Altman (1966), and Schmuck and Schmuck (1971).

McGrath and Altman (1966) suggested ten variables all known to influence the output of any problem solving group. These variables include: (1) member abilities and experience, (2) member attitudes, (3) member roles and/or tasks, (4) group size, (5) group task, (6) group leadership, (7) group developmental stage, (8) group cohesiveness, (9) environmental variables, and (10) group organization and/or structure. An attempt is made to deal with most of these variables in development of the MK II debriefing procedures.

Group organization and structure.--Maier (1963) described several techniques or strategies for organizing group problem solving activity which may be dichotomized as structured or unstructured. Since unstructured strategies normally take more time, they were not considered appropriate for the present study.

Among the structured techniques for organizing a small problem solving group, the most applicable appeared to be "problem posting" (Maier, 1963, p. 161). Using this technique, the student participants are given a common experience, e.g., individual use of the prototype SLATE materials. Following this, they convene for a debriefing. The first part of the debriefing, however, is devoted to listing all the problems encountered by various members of the group. During this

time the group leader summarizes the problems and writes them on a blackboard--thus collecting data and assisting the group and himself to conceptualize the problems encountered by various individuals in the group.

The list of problems is then made the subject of an organized discussion in which the group assumes responsibility for development of solutions to each problem. When time does not permit an exploration of all problems, the group is allowed to select those of greatest interest. Maier cited evidence that this technique is effective in stimulating interest, helps problem conceptualization, and leads to greater productivity in developing solutions (Maier, 1963, p. 191).

In light of the foregoing discussion, it was determined by E that a small group problem posting debriefing would be the best format for generating the types of data required to revise prototype SLATEs.

Group leadership.--Most of the research information about leadership performance came from studies of leaderless group situations, although some data came from studies using superiors' ratings of leadership in operational settings (McGrath & Altman, 1966). Effective leadership has been shown to be a function of a number of characteristics and conditions including education, intelligence and/or task ability, high group status, training in leadership techniques, communication skills, and individual personality characteristics such as extroversion, assertiveness and maturity.

In the present study, it was determined that the prototype lesson author would be designated the group discussion leader by virtue of his expertise in the subject matter and his responsibility as the instructor

in the course. Assuming that the personality characteristics, education, communication skills, and intelligence of lesson authors (group leaders) cannot be changed, some benefit might accrue through training in group leadership techniques. However, due to lack of faculty time it was felt that any systematic group leadership training program for SLATE designers was out of the question. Therefore, as an alternative, a "debriefing checklist" was developed by E which outlined the ground rules, tasks, and responsibilities of all participants (Appendix D).

Group size.--The size of the group was determined largely by research on group processes and logistic considerations. For example, Maier (1963) cited evidence that greatest productivity in problem solving groups is often obtained when the group contains between six and ten participants. Logistically, six to ten students from the target population should be readily available when the opportune time for tryout is reached. The optimal size decided upon was nine students plus the group leader (SLATE author) for a total of ten participants.

Group composition.--The composition of the group was determined by the desire to obtain a sample which represented as nearly as possible the spectrum of abilities in the target population. It was assumed that students of different entering abilities but similar prerequisite knowledge would encounter different learning problems with prototype SLATEs, and it would be valuable for the SLATE author to be confronted with these problems. Furthermore, it was hoped that by varying the group composition between high and low ability students, the high ability students could assist the SLATE author in determining solutions to problems encountered by themselves and the low ability students. It was possible

that the opposite might also occur; e.g., low ability students could assist in solving high ability students' problems.

Assuming the desirability of using students of varying ability in the group, the Scholastic Aptitude Test (SAT) was selected as a normalized measure of entering students' abilities. This measure was selected mainly because SAT scores on most students in the target population (Michigan State University) were already available. It was felt that SAT was equally as valid as other measures for purposes of selecting students possessing a range of abilities. For other target populations, other normalized ability measures might be selected. It is the experimenter's opinion that the choice of a specific measure of ability is not as important as the procedure of using a normalized measure to select students possessing a range of abilities.

Group and individual tasks.--The task of the group as designated in the ground rules was to provide the group leader information regarding identification and remediation of instructional problems.

The general orientation given the students was to participate in lesson development as co-authors (Yelon & Scott, 1969). That is, the students were asked to share the responsibility for providing data on their learning problems as well as suggest solutions to these problems. The task of student participants was twofold. First was individual student interaction with the prototype SLATE materials. For logistic simplicity, students were requested to use the materials within some reasonable time period. After allowing some time for scoring lesson evaluation instruments, the debriefing began to take advantage of immediate reminiscence of learning problems.

The task orientation and preparation given the group leader (SLATE author) was: (1) to study and use the "debriefing procedures" checklist; (2) to adopt an attitude that "the materials are on trial, not the students:" and (3) commitment to the principle of "no reprisals" for frank and/or derogatory comments.

The leader's task during students' use of the materials was to be that of a tutor offering assistance as required to individual students. As a student indicated a problem, the SLATE author was to visit the student, note the problem and its location in the SLATE, answer the student's question, and discuss these problems during the debriefing. Presumably, if a number of students (30% or more) had similar problems, a revision was to be made so that the actual tutorial instruction would be incorporated into the SLATE.

During the debriefing, the SLATE author should function as a data collector posting the problems and organizing the data so that later discussion could focus on solutions to the problems posted. Obviously, different authors would vary in their group interaction skills, and these differences would affect the quantity and quality of data collected. Nevertheless, direct face-to-face confrontation with learning problems provides an experiential dimension which is likely to convince authors that certain problems must be remediated.

The time limits of the total group process were established arbitrarily after consultation with several potential participants in the field trial part of the study. These authors indicated they would not participate in obtaining feedback from students any longer than two or three hours maximum--per SLATE. Therefore, a two hour limit was established for the group debriefing process.

Student experience and attitudes.--To obtain valid information on instructional problems, students would necessarily be selected from the target population for whom the prototype SLATE was intended. Students should possess necessary SLATE prerequisites but not score higher than the chance level on the lesson pre-test.

To ensure some degree of success in obtaining the desired feedback, students should possess a positive attitude towards the task of the group. Selection of students from a pool of volunteers is assumed to meet the requirements of obtaining students with a positive task orientation.

Summary of the Group Debriefing Technique Incorporated in the MK II Model

A group process methodology was substituted for both the tutorial and large group tryout procedures specified in the MK I model, thus overcoming many of the objections cited by respondents. The overt objectives of the group process were twofold: (1) to generate data on SLATE deficiencies/instructional problems, and (2) to develop feasible solutions to these problems. A covert or "hidden agenda" objective was to provide the SLATE author an opportunity to observe personally the deficiencies in the prototype and thus help overcome the natural reluctance to revise.

The group process methodology is shown in Figure 6 and essentially involves the following components: (1) selection of nine volunteer students who vary in their entering abilities (SAT scores), (2) individual use of the prototype SLATE materials by these volunteers, (3) administration and assessment of learning and attitudinal measures to provide a basis for conducting an organized debriefing, and (4) participation in a group debriefing following use of the materials which involves problem posting and problem solving techniques.

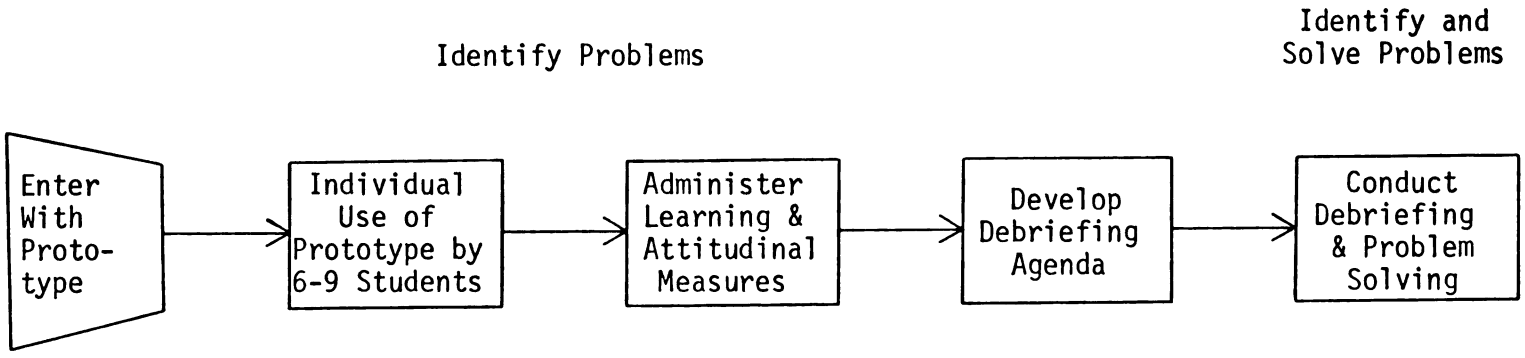


Figure 6.--MK II Group Debriefing/Problem Solving Technique

Description of the MK II "Mini" and "Maxi" Models

Previous discussion has presented the rationale for major modifications to the MK I model. In revising the MK I model it was deemed necessary to create two revised versions which are designated the "mini" and "maxi" MK II models. The "mini" version is highly simplified in order to facilitate conceptual understanding of the process. The "maxi" version is highly detailed and intended for use by consultants or with faculty who are intimately familiar with the "mini" version.

MK II "Mini" Model

Basically, the MK II "mini" model is a flowchart specifying the chronological sequence of tasks which are to be performed by an author during formative evaluation of his SLATE (see Figure 7). Each task contributes to a function essential to the total process. In all, there are five basic functions: (1) logistics, (2) data collection, (3) data analysis, (4) revision design, and (5) recycle.

At least two iterations are required to complete the process because the model stipulates that data be collected from two fundamentally different sources of information and revisions be developed sequentially based on these two sources of feedback. These sources of information are: (1) technical consultants and (2) volunteer students. Each source provides feedback on basically different types of problems. Technical experts, for example, provide feedback on discrepancies in subject matter content, instructional media, and in evaluation instrument design. Volunteer students, on the other hand, function to provide feedback on their specific learning problems. Both sources complement each other so that the widest range of discrepancies can be identified in a minimum length

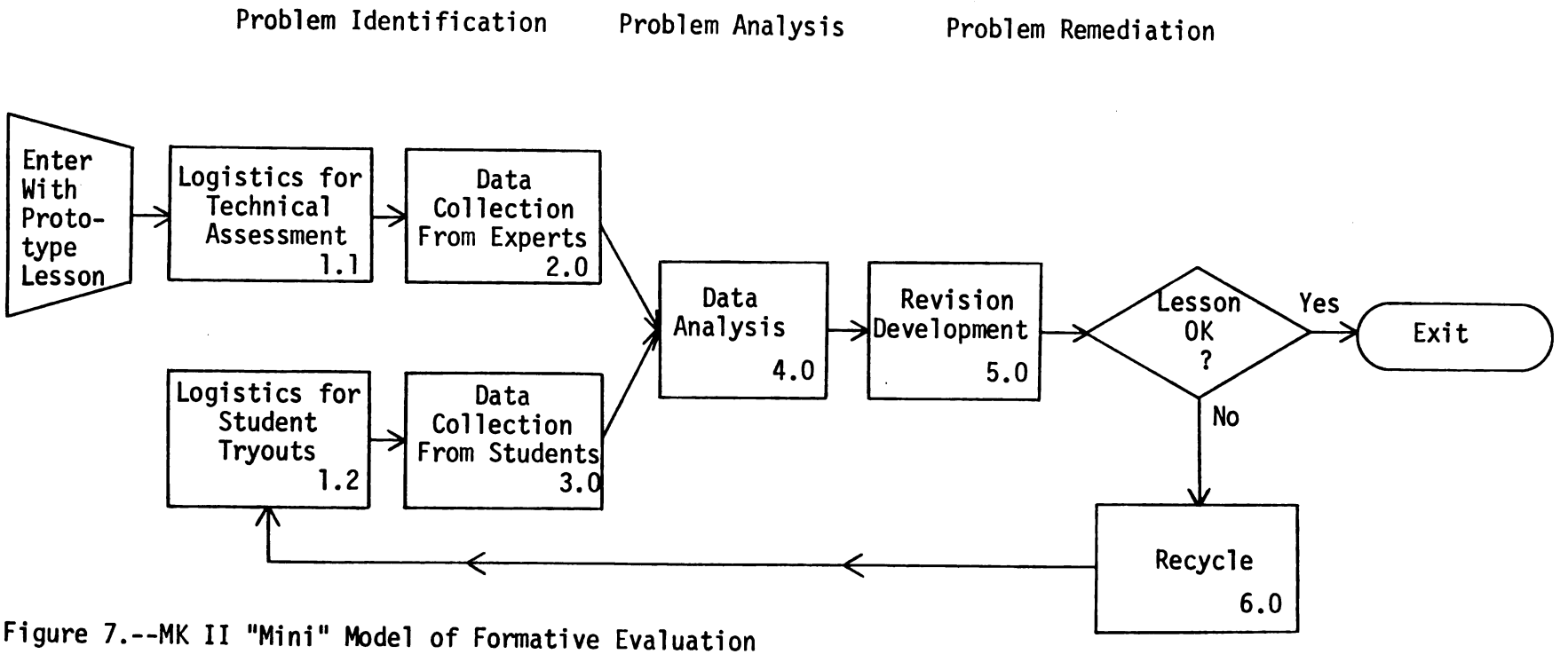


Figure 7.--MK II "Mini" Model of Formative Evaluation

of time. The process begins when a prototype instructional unit is completed to the point where the author believes it is ready to be used with students.

Assuming "readiness" of a prototype SLATE, the MK II formative evaluation process consists of two cycles of "problem identification," "problem analysis," and "problem remediation." In the first cycle, technical problems are identified by feedback from technical experts who review the new instructional unit. Following collection of data on technical discrepancies, the SLATE author analyzes these problems in conjunction with an instructional development or learning specialist, and revisions are developed. The process then recycles so that in the second cycle, learning problems are identified through feedback from a group of volunteer students from the target population. Again, following collection of these data, the SLATE author analyzes the problem with an appropriate consultant and revisions are developed.

It is important that technical discrepancies be remediated before student tryout of the prototype SLATE. The reason for this sequence is that SLATE authors vary considerably in their media design and production skills, their knowledge of and ability to organize subject matter, and in their skill in designing evaluation instruments appropriate to formative evaluation. To preclude students' learning erroneous content, being confronted with illegible or inaudible stimuli, and/or avoidance of critical omissions in evaluation instruments, the SLATE author must obtain feedback from the technical experts and revise the prototype prior to student tryouts.

The number of cycles required to bring the prototype up to operational readiness would vary depending on how "bad" the prototype was and

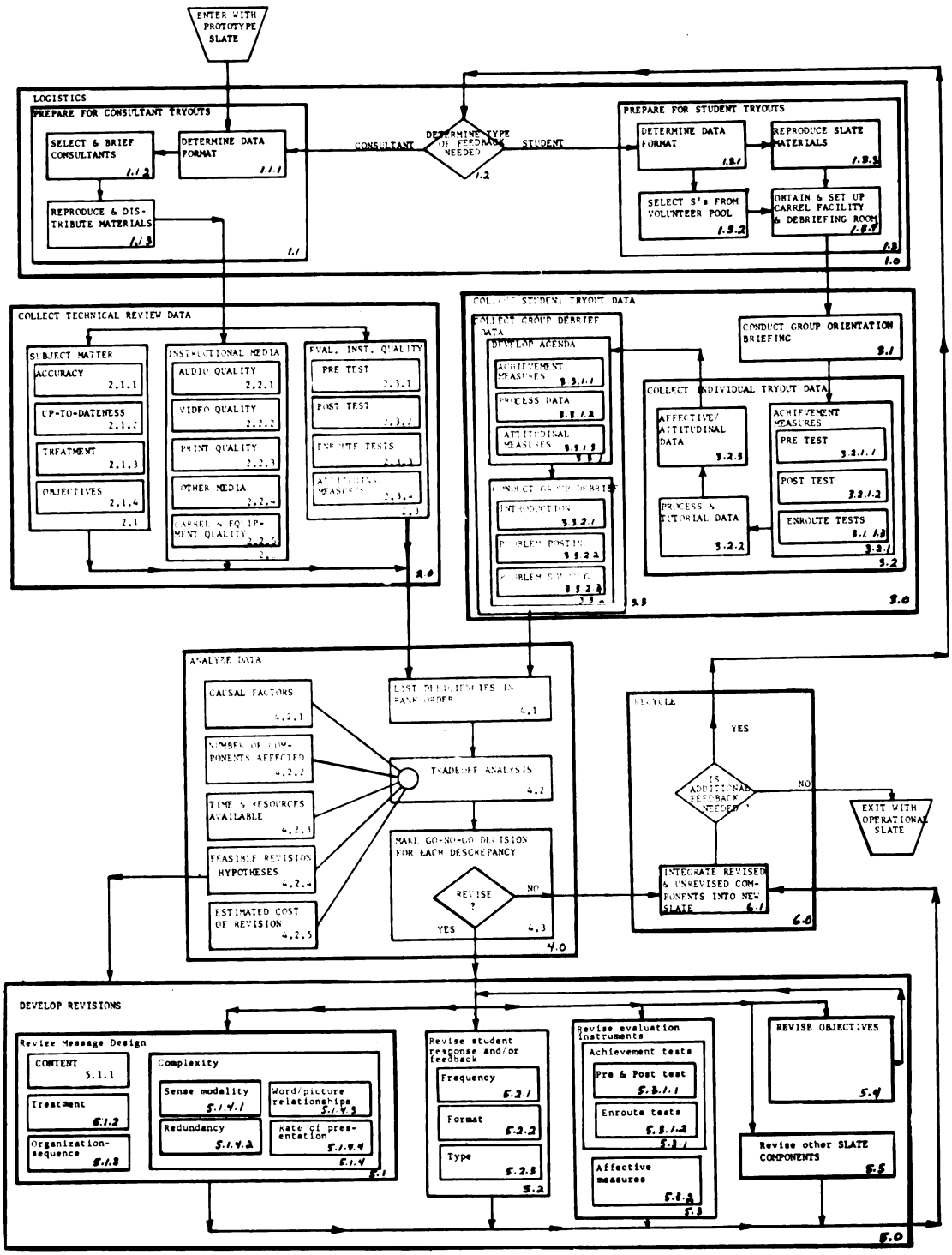
how stringent the operationally ready criteria are. In the present study "operational readiness" was defined as: (1) 80% or more of the student tryout group achieving 80% or higher on the post-test, and (2) not more than 20% "unsatisfactory" responses on the post-instruction attitude survey (Appendix G).

In sum, the purpose of the MK II "mini" model is to provide a framework to familiarize authors with the process of identification and remediation of problems which interfere with student achievement of intended learning objectives. The MK II model basically consists of two cycles of five similar functions; the major difference between cycles being the source of information which provided the feedback during data collection. Cycle one essentially serves a technical quality control function by obtaining feedback from technical experts and developing revisions based on this feedback. Cycle two serves to remediate specific student learning problems by obtaining feedback from volunteer students in the target population and devising revisions to alleviate these discrepancies.

The two cycles are complementary in that through use of two different sources of feedback, the widest range of discrepancies can be identified and remediated.

MK II "Maxi" Model

The MK II "mini" version just described serves a useful purpose in orienting users to the process of tryout and revision. After orientation however, there is a need for detailed instruction and specification of techniques for carrying out the process. This detail is provided in the MK II "maxi" model, shown in Figure 8. Because the reader need


ENTER WITH PROTOTYPE SLATE
LOGISTICS
PREPARE FOR CONSULTANT TRYOUTS
SELECT & BRIEF CONSULTANTS 1.1.2
DETERMINE DATA FORMAT 1.1.1
REPRODUCE & DISTRIBUTE MATERIALS 1.1.3
1.1
CONSULTANT
DETERMINE TYPE OF FEEDBACK NEEDED 1.2
STUDENT
PREPARE FOR STUDENT TRYOUTS
DETERMINE DATA FORMAT 1.3.1
REPRODUCE SLATE MATERIALS 1.3.3
SELECT S's FROM VOLUNTEER POOL 1.3.2
OBTAIN & SET UP CARREL FACILITY & DEBRIEFING ROOM 1.3.4
1.3
1.0
COLLECT TECHNICAL REVIEW DATA
SUBJECT MATTER
ACCURACY 2.1.1
UP-TO-DATENESS 2.1.2
TREATMENT 2.1.3
OBJECTIVES 2.1.4
2.1
INSTRUCTIONAL MEDIA
AUDIO QUALITY 2.2.1
VIDEO QUALITY 2.2.2
PRINT QUALITY 2.2.3
OTHER MEDIA 2.2.4
CARREL & EQUIPMENT QUALITY 2.2.5
2.2
EVAL. INST. QUALITY
PRE TEST 2.3.1
POST TEST 2.3.2
ENROUTE TESTS 2.3.3
2.3
2.0
COLLECT STUDENT TRYOUT DATA
COLLECT GROUP DEBRIEF DATA
DEVELOP AGENDA
ACHIEVEMENT MEASURES 3.3.1.1
PROCESS DATA 3.3.1.2
ATTITUDINAL MEASURES 3.3.1.3
3.3.1
CONDUCT GROUP DEBRIEF
INTRODUCTION 3.3.2.1
PROBLEM POSITING 3.3.2.2
PROBLEM SOLVING 3.3.2.3
3.3.2
3.3
CONDUCT GROUP ORIENTATION BRIEFING 3.1
COLLECT INDIVIDUAL TRYOUT DATA
AFFECTIVE/ ATTITUDINAL DATA 3.2.3
ACHIEVEMENT MEASURES
PRE TEST 3.2.1.1
POST TEST 3.2.1.2
ENROUTE TESTS 3.2.1.3
3.2.1
PROCESS & TUTORIAL DATA 3.2.2
3.2
3.0
ANALYZE DATA
CAUSAL FACTORS 4.2.1
NUMBER OF COMPONENTS AFFECTED 4.2.2
TIME & RESOURCES AVAILABLE 4.2.3
FEASIBLE REVISION HYPOTHESES 4.2.4
ESTIMATED COST OF REVISION 4.2.5
LIST DEFICIENCIES IN RANK ORDER 4.1
TRADEOFF ANALYSIS 4.2
MAKE GO-NO-GO DECISION FOR EACH DISCREPANCY
REVISE ? 4.3
NO
YES
4.0
RECYCLE
YES
IS ADDITIONAL FEEDBACK NEEDED ?
NO
EXIT WITH OPERATIONAL SLATE
INTEGRATE REVISED & UNREVISED COMPONENTS INTO NEW SLATE 6.1
6.0
DEVELOP REVISIONS
Revise Message Design
CONTENT 5.1.1
Treatment 5.1.2
Organization-sequence 5.1.3
Complexity
Sense modality 5.1.4.1
Word/picture relationships 5.1.4.3
Redundancy 5.1.4.2
Rate of presentation 5.1.4.4
5.1.4
5.1
Revise student response and/or feedback
Frequency 5.2.1
Format 5.2.2
Type 5.2.3
5.2
Revise evaluation instruments
Achievement tests
Pre & Post test 5.3.1.1
Enroute tests 5.3.1.2
5.3.1
Affective measures 5.3.2
5.3
REVISE OBJECTIVES 5.4
Revise other SLATE COMPONENTS 5.5
5.0

not have intimate knowledge of MK II "maxi" procedures to understand the thrust of the present study, the detailed explanation of MK II "maxi" is placed in Appendix B.

Chapter Summary

This chapter has described the first phase of validation of the model of formative evaluation being developed in the present study. This first phase of validation consisted of interviewing seven selected SLATE authors to determine their opinions on the practicability of the MK I model and assess the degree to which MK I is congruent with their personal tryout and revision procedures.

Interviews were conducted with six non-professional and one professional SLATE authors. The net result of these interviews was recognition on the part of the experimenter that the MK I model differed considerably from current practice of the respondents. In general, the respondents felt the MK I model was impractical. The major problems with MK I were that it appeared too time consuming, costly, and logistically difficult to integrate into ongoing teaching activities.

As a result of these data, major modifications were made to the MK I model, resulting in MK II "mini" and "maxi" models. The major difference between MK I and MK II versions is the inclusion in the MK II of a student group debriefing. This debriefing follows student use of the prototype instructional stimuli and is organized to follow a problem posting and problem solving format. It is reasoned that by means of the debriefing procedure students can aid the SLATE author in developing solutions to the problems identified in a minimal amount of time.

Following development of the MK II version, the study progresses to the second stage of validation: field test of the MK II version with three Michigan State University SLATE authors. The methodology for the field tests is outlined in Chapter IV, and the field tests themselves are described in Chapter V.

CHAPTER IV

METHODS AND PROCEDURES

The research methods and procedures used in five field trials to investigate the efficacy of the MK II model are described in this chapter.

Two distinct types of research objectives were being sought in this study. The first was related to experimentally comparing student achievement and attitudes resulting from a prototype (unrevised) SLATE with the revised counterpart.

The second type of objective centered on understanding and describing the process through which the experimental treatments came into being. In this study, the experimental treatments (revised SLATEs) were developed on the basis of procedures in the MK II model. Since the MK II model was itself a prototype, a description of the problems and successes resulting from its procedures was essential for further modification and refinement of the model.

Research Strategy

The overall research strategy called for five field experiments in three disciplines to include gathering and analysis of both descriptive and experimental data. Essentially each field experiment represented a replication of the developmental process, that is, application of the MK II model in a field setting. It was felt that five replications involving three different authors and academic disciplines would provide a sufficient number of trials to: (1) identify critical variables in the

process, (2) suggest modifications to the model, and (3) establish the validity, feasibility, and effectiveness of the MK II model.

Descriptive Methodology

Data Collection

Descriptive data were collected using the basic technique known as high inference observation (Kerlinger, 1964, p. 510). Using this method, an observer abstracts relevant information from his ongoing observations and later makes inferences about variables.

The experimenter (E) had the dual responsibility of interacting with each author (Author A, B, and C) on a consultant basis, as well as observing and recording the nature of these interactions and subsequent decisions. Narrative data were collected at each meeting between experimenter (E) and individual SLATE authors (A, B, and C). During these meetings, E kept a "log" which was then summarized and combined with impressionistic data in a memorandum written immediately following each meeting. Inferences, problems, and suggestions were included in the last section of each memorandum. Tape recordings supplemented E note taking during the very critical author-student feedback interactions at control group and experimental group tryouts. But all other descriptive data were gathered by E observation and note taking.

At the conclusion of each field experiment, the memos from each meeting were summarized to form a narrative description of the whole formative evaluation/developmental process. These narrative descriptions were systematically related to procedures in the model and reported in Chapter V, "Description and Results of Five Field Trials."

Experimental Procedures and Methodology

Similar procedures and methodology were used to conduct experimental comparisons between original and revised SLATEs in three field experiments, A_1, A_2, and B_1. In two field experiments, A_3 and C_1, experimental treatments were not developed. Therefore, the following description of experimental procedures apply only to A_1, A_2, and B_1.

Experimental Design

The basic experimental design used in this study was the before-after control group design (Campbell & Stanley, 1963) illustrated in Figure 9. This design has been criticized by Kerlinger (1964, p. 310) for its use of pre-tests which may be reactive. That is, experimental Ss may become sensitized to the criterion test items and may then be responding to a combination of reminiscence of test items as well as the experimental treatment. In the present study, this sensitization effect was not considered a problem, but, quite the contrary, as an advantage. Pre-test items were regarded as "advanced organizers" and operational definitons of SLATE objectives. Sensitization to objectives by means of test-like events may enhance learning (Rothkopf, 1966, 1968) so pre-tests were considered essential and integral parts of both experimental and control group treatments.

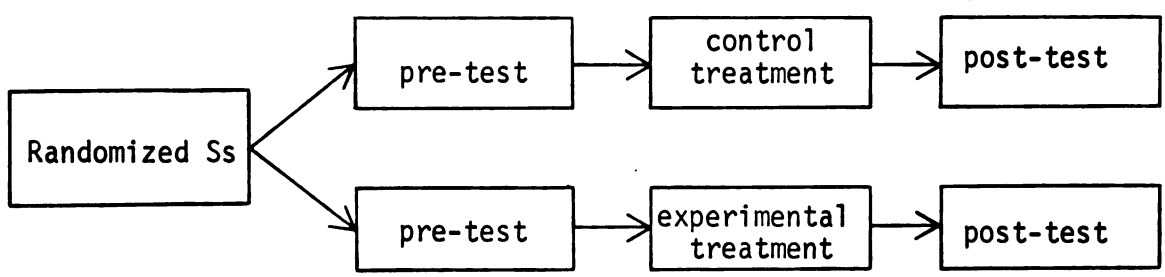


Figure 9.--Before and After Control Group Design

Selection of SLATE Authors

The three participating authors (A, B, and C) were selected on the following bases:

1. They were currently teaching a course using SLATEs which they had personally developed.
2. They had developed a prototype SLATE for use in their course which had not previously been used by students or undergone any formative evaluation.
3. They were willing to participate in this study with the understanding that volunteer students from their current course would provide feedback on their prototype SLATE; the total process of data gathering and revision would likely take 20-30 hours of their time; it was likely, but not certain, that the revised SLATE would be better than the original.
4. They had similar backgrounds and experience in programmed instruction and SLATE design, but were from different academic disciplines.

Author A participated in formative evaluation of three SLATEs, designated A_1, A_2, and A_3. Authors B and C each conducted formative evaluation of one SLATE, designated B_1 and C_1.

Additional background information on Authors A, B, and C is contained in Appendix F.

Selection of Students

Population.--The populations from which students (Ss) were selected were defined as the target populations for which the prototype SLATEs were intended. Three populations were involved; specifically, the students enrolled in three courses at Michigan State University, Fall term, 1970. These three courses were: (1) Animal Husbandry 111 (an introductory course for majors); (2) Education 327M (an introductory course for teachers of secondary school industrial arts, metalworking); and (3) Biology 141 (an introductory course in biology for majors). These courses were taught by the three participating SLATE authors.

Stratified random sampling.--Sampling procedures treated Ss from each course as essentially different populations due to differences in subject matter content and prerequisite skills involved. Selection of Ss for experimental and control groups was predicated on four criteria: (1) voluntary status, (2) stratification by SAT score, (3) randomization, and (4) Ss would possess prerequisite skills required by the prototype SLATE, but would be naive with respect to the terminal objectives.

After consultation with SLATE authors, agreement was reached as to the most appropriate time in the course sequence to run the experiment. Authors agreed to withhold information in their courses which might bias Ss until after the control and experimental groups had been conducted.

About one week prior to prototype (control group) tryout, authors personally solicited volunteers from their classes. The experiment was described as a learning experience in which all class members would have to participate eventually, but that some volunteers were needed immediately to provide constructive feedback on a prototype version. This feedback would be used by the author to revise the SLATE and hence improve the learning experience for those to follow. Solicitation was successful in that a sufficient number of volunteers were obtained to permit stratification and randomized assignment to treatments.

After obtaining a pool of volunteer Ss from each population, E obtained Scholastic Aptitude Test (SAT) scores from University records. Volunteers not having SAT scores were dropped from the pool.

Within the volunteer pool from each class, E stratified Ss into High, Medium, and Low sub-groups. This was done by making a rank order list by SAT, for each pool of volunteers, then partitioning each ranking

into thirds, for three sub-groups. Ss from each sub-group were selected randomly and alternately assigned to control or experimental groups until each treatment had an N=12 consisting of four high, four medium, and four low SAT Ss. A schematic of the sampling procedure used for the three experimental comparisons is shown in Figure 10.

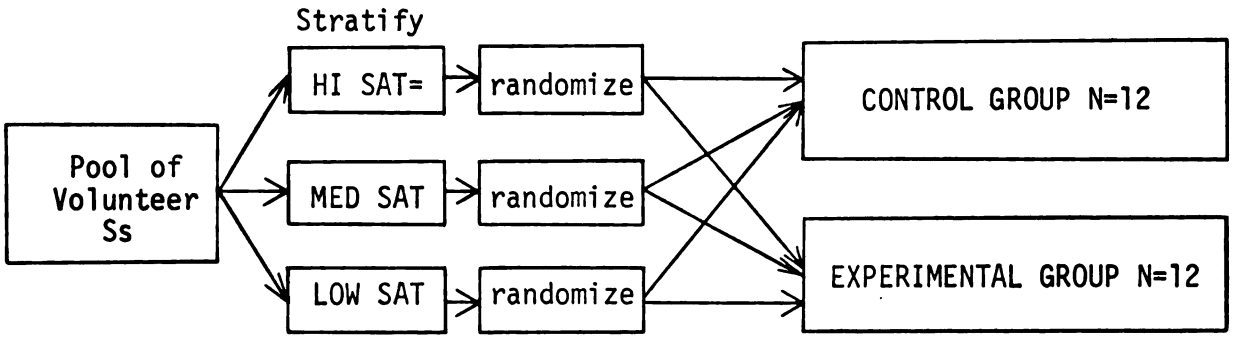


Figure 10.--Procedure for Assignment of Ss to Treatments

In one case (B_1), however, so much time elapsed between control group tryout and development of the revised version (seven months), that Ss originally designated for the experimental group were no longer naive with respect to the content of the prototype SLATE. Consequently, a second call for volunteers was made from an equivalent population (same course, two terms later) and stratification and randomization techniques were used to select the experimental group.

In all three experimental comparisons, Ss were volunteers from the ongoing course, SAT scores were used as the partitioning variable, equal numbers of Ss from high, medium, and low sub-groups were represented in experimental and control treatments, and pre-experimental equivalence was substantiated by comparison of pre-test scores between experimental and control groups.

Treatments

The five prototype SLATEs used in the study were all to be used in ongoing courses at Michigan State University, Fall term, 1970. Author A developed three prototype SLATEs, designated A_1, A_2, and A_3, to be used in his undergraduate service course in Animal Husbandry (AH 111). These SLATEs were entitled "Pork Carcass Evaluation," "Cattle Breeds," and "Cattle Carcass Evaluation." This course enrolls 175 students per term, primarily freshman and sophomores, who are heterogeneous in terms of major fields, motivation, and background.

The instructional method used in the course consists of two lectures, two SLATEs and one laboratory per week. Students would therefore be very familiar with the SLATE self-instructional environment.

The fourth SLATE, developed by Author B (designated B_1), was a lesson on "How to Read and Care for a Micrometer." This SLATE was to be used in an undergraduate course in industrial arts enrolling fifty industrial arts majors, primarily juniors and seniors. No other SLATEs are used in this course, so students were not familiar with the format.

The fifth SLATE developed by Author C (designated C_1) was to be used in a freshman biology course serving 150 majors in a residential college. The prototype SLATE used in this study was an overview of several types of ecological systems, entitled "The Schema Biologica." Since other SLATEs were used in this course, students would be familiar with this technique.

Control treatment.--All control group treatments involved Ss' use of unrevised prototype SLATE materials which had been reviewed by E for evaluation instrument quality and reviewed by author peers for

content accuracy. Control treatment SLATEs consisted of pictorial information on 35 MM slides and in student workbooks, audio information on a tape recording, printed information in the student workbook, pre- and post-tests and a post instruction attitude survey. In A_1 and A_2, Ss responded to these materials individually in learning carrels. Students thus proceeded at their own rate, controlling number of repetitions of slides and tapes, and response rate in their workbooks. (Any time they repeated slides or tape, they were asked to note this activity.) Audio information was presented via headphones, and Ss were asked not to interact with one another but to direct any questions to the SLATE author who was available in the carrel room.

In B_1, however, insufficient carrels were available for simultaneous individual student participation prerequisite to the group debriefing. Therefore, out of necessity, a group presentation mode was adopted instead of individual presentations. In the group mode, the SLATE author controlled a single slide projector and tape recorder, stopping or repeating the presentation at the request of any S. Ss' responses were, nevertheless, still recorded individually in their workbooks. When a S stopped the presentation by asking a question, obviously the whole group was affected. Consequently, B_1 was not really a close simulation of a self-instructional environment. Nevertheless, since the purpose of the group was to provide feedback to the SLATE author, the technique was considered valid. However, caution must be used in interpreting scores on post-tests as these scores are likely to be inflated as a result of group discussions during the original presentation.

Experimental treatments.--Each experimental treatment consisted of Ss using the revised set of slides, audio tape, workbook, pre- and post-test, with the attitude survey unchanged. A_1 and A_2 again used the self-paced carrel mode and B_1 used the group presentation mode. In two cases (A_1 and A_2), the elapsed running time (no playbacks) of the revised versions was reduced 20%; on the other hand, B_1 elapsed time was increased 50% (17 minutes to 26 minutes). Development of experimental treatments are reviewed in the next chapter.

Independent Variable

The independent variable in each experimental comparison was conceptualized as the total set of procedures, operations, and decision rules contained in the MK II model of the formative evaluation process (Figure 8), plus unique contributions by the users of the model (E and SLATE author). In short, the independent variable was the model and its application.

Dependent Variables

Four dependent variables were used as criteria for assessing the effect of the independent variable.

1. Group Mean Achievement.--Intended as an immediate post measure of student achievement of terminal objectives.

2. Gain Score.--Mean difference between pre-test and post-test scores.

3. Percentage of Students Achieving "Mastery."--Intended as a criterion referenced measure to determine which treatment enabled a greater number of Ss to achieve a minimum acceptable level of performance, e.g., 80% or more correct on post-test.

4. Student Attitudes.--Intended as an immediate post measure of student perceptions of lesson deficiencies and strengths.

Development of Instruments

Generally, two types of instruments were developed. First, measures of student achievement specific to a given SLATE were developed by each SLATE author in consultation with E. Second, a Likert-type instrument was developed by E to assess student perceptions of lesson strengths and weaknesses.

Achievement measures.--Student learning on each SLATE was measured by SLATE author designed pre- and post-tests. Pre-tests contained a caveat to reduce anxiety or frustrations resulting from a low score, but cautioned Ss that 80% criterion was required on the post-test. The post-test and pre-tests used identical items and a self-scoring format. This format was selected because additional learning would likely occur as Ss scored their tests.

The overriding majority of these particular SLATE objectives were cognitive; e.g., recall, visual or verbal discrimination, or problem solving. B_1, however, did require an integration of cognitive and perceptual motor skills (measurement with a micrometer). In light of the preponderance of cognitive objectives, achievement measures were largely paper and pencil variety. At E's suggestion, item forms were deliberately varied to include true-false, multiple choice, completion, and matching.

During initial review of these achievement tests, E noted a number of discrepancies in that test items did not reflect stated SLATE objectives. This problem was compounded by the fact that in no case were SLATE objectives stated in behavioral terms. Consequently, E consulted with

each author approximately four hours per SLATE helping operationalize their objectives and translate these operations into test items.

As finally developed, pre- and post-tests included many items in common with the en route self-tests. Particular attention was given to articulating post-test items with en route self-tests so errors on the post-test could be linked back to that place in the SLATE where instruction was accomplished.

Feedback from students (control group) showed that numerous items on the prototype achievement measures were faulty. These items were then either deleted completely, thus reducing the total number of items, or were replaced by new and presumably better items. Thus, experimental and control group achievement measures were not totally identical.

To assess the statistical significance of differences between experimental and control achievement measures, only those items common to both original and revised measures were used. The total number of items on original and revised measures and number of items common to both is shown in Table 6.

Table 6.--Number of Items on Pre- and Post-Tests

		Total Items on Pre- and Post-Test	Total Items Common Between Experimental and Control Group
A_1	Control	60	40 items worth 47 points possible
	Experimental	52	
A_2	Control	56	40 items worth 40 points possible
	Experimental	47	
B_1	Control	15	15 items worth 15 points possible
	Experimental	15	

Scoring and data display.--All pre- and post-tests were self-scored by Ss. To reduce cheating, Ss were given an answer key when nearly finished with each test. Furthermore, before data analysis was begun, all scores (totals and individual items) were rechecked by E for accuracy. (E noted about 10% scoring error rate, usually with the error raising the S's score.)

During control and experimental group tryouts, test scores were displayed on an item by student matrix (Appendix E). This method enabled E and SLATE author to identify items missed by over 30% of the group and any such item became a topic of discussion at the group debriefing.

Attitudinal measure.--A post instruction attitude survey was developed by E specifically to measure Ss' perceptions regarding several aspects of the SLATE they had just finished (Appendix G). Specifically, this instrument was a twenty-seven item Likert-type rating scale seeking to measure four general factors.

1. SLATE strengths and weaknesses resulting from communication/message design factors:

Factor	Item Number
a. Rate of presentation	8
b. Redundancy	9
c. Interest and attention	5
d. Clarity of instruction and examples	11, 13, 15
e. Vocabulary level	16
f. Audio and video quality	7

2. SLATE strengths and weaknesses resulting from learning or task factors:

a. Prerequisites	1
b. Objectives	2
c. Motivation	3
d. Organization and sequence	6, 14
e. Evaluation and feedback	17, 18
f. Type of response and frequency	12, 19
g. Relevancy of information	10

3. SLATE strengths and weaknesses resulting from management/ technical factors:

Factor	Item Number
a. Equipment manipulation	4
b. SLATE methodology	28
c. Tryout procedures	27
d. Degree of revision needed	22

4. Perceived learning and attitudes resulting from the lesson:

a. Attitude towards subject matter	30
b. Terminal understanding of concepts	26
c. En route understanding of concepts	29
d. Certainty of learning	20
e. Amount of learning	21

In addition, four open-ended questions were included to encourage students to express opinions and perceptions not previously accounted for in the Likert items.

The attitude survey instrument was used in all experimental and control groups. Few criticisms of this instrument were obtained during debriefings; hence items were not modified and the rating scale as originally drafted was used throughout.

Scoring and data display.--During each experimental and control group, the attitude survey was scored by E immediately after completion by each S. A numerical value from one to five points was assigned to each response, five representing the "ideal" response and one representing a very low or dissatisfaction response. Total scores for individual Ss were tallied, but more important, a running tally was kept for each item on the attitude survey. If a S's response deviated by more than two points from the ideal, it was tallied. Each item, which 30% or more of Ss had rated too far from ideal became topics of discussion at the debriefing. In addition, if 30% of the "open-ended" responses were on a similar topic, that topic was discussed during the debriefing.

Experimental Procedures

A_1 and A_2 used identical procedures; however, B_1 varied in several respects and will be described separately.

A_1 and A_2 procedures.--After experimental and control groups were selected, E coordinated scheduling of the SLATE author, the carrel facilities, and the Ss by selecting a date and time for the experiments and asking Ss to RSVP, regrets only. Ss who had a scheduling conflict were traded among treatments, provided they were in the same SAT sub-group. If no trade-off was possible, the originally selected S was dropped and another selected from the pool of volunteers, within the given SAT sub-group.

Data collection in both A_1 and A_2 experimental and control groups were conducted from 7:00 until 10:00 p.m. in the carrel facility in the Department of Animal Husbandry, 108 Anthony Hall, Michigan State University, during Fall term, 1970. This facility can accommodate twelve individual students maximum.

To reduce possible bias from Ss' social interaction, the experimental treatment was developed and administered as rapidly as possible following the control group data gathering. In A_1 and A_2 the time interval between administration of control and experimental treatments was one week.

E developed an "agenda" for the conduct of the experimental and control treatments which was discussed extensively with the participating author several days prior to the first tryout (Appendix C). After the discussion, E provided the SLATE author with a checklist to guide the treatment activities. It was determined that the SLATE author rather than the experimenter should conduct the experiment, in the sense of

providing instructions to the Ss, answering their questions, and conducting the debriefing. E would be present to observe the process, collate and score instruments, and remediate minor technical difficulties; but operationally, each treatment was conducted by the SLATE author. (This decision was made to see if the procedures in the agenda could be carried out competently by the SLATE author; if not, what changes would have to be made so the procedure would be independent of E.)

Since the complete agenda and checklists are included in the appendix, they are not reiterated here. Instead, a narrative summary of the procedures are presented.

On the evening of a treatment, SLATE author and E arrived one hour early to inspect all carrels to prevent obvious technical malfunctions such as inoperative or missing equipment or slides improperly positioned. As Ss arrived, name tags were provided and SLATE author began non-course related "small talk" to place Ss at ease. After all Ss had arrived, E tape-recorded the remainder of the session. The formal treatment began with a 10-15 minute orientation briefing by the SLATE author designed to do the following:

1. Express appreciation for Ss' participation and orient Ss as to the purpose of the session.

2. Relieve Ss' anxiety and facilitate their open and frank interaction.

3. Describe the planned sequence of events which were:

    a. Pre-test
    b. Individual use of treatment AV materials
    c. Post-test
    d. Attitudinal survey
    e. 15-minute "break" including refreshments
    f. Reconvene for debriefing and feedback session

4. Establish the "ground rules" for the session which were:

   a. No talking to each other during lesson
   b. Take notes on type and locating of problems; e.g., don't understand, bored, lesson too fast, etc.
   c. Raise hand for tutorial assistance
   d. Score own pre- and post-tests
   e. Do not cheat
   f. Do not discuss SLATE during the break
   g. Please remain for the debriefing

It was repeatedly emphasized that in no way would Ss' remarks be used in a punitive sense.

Following the orientation briefing, Ss selected a carrel and worked on the pre-test. As they neared completion, E distributed pre-test answer sheets. Ss were allowed to begin the lesson immediately after completing and scoring the pre-test. There was usually a 5-10 minute differential among Ss regarding pre-test completion time.

When all Ss were working on the lesson, E collected all pre-tests and answer sheets. Ss' scores were rechecked and placed on an item-student matrix for display. In most cases Ss achieved below chance level, although one or two scored 70% correct. (Later discussion with these Ss showed they were guessing.)

While Ss interacted with lesson materials, Author A circulated freely answering questions on a tutorial basis and made notes of the questions and his responses. All such interactions were tape-recorded by E.

Post-tests were distributed as Ss neared completion of the SLATE. Answer keys and attitude surveys were distributed as Ss neared completion of the post-test. Ss returned scored post-tests and unscored attitude survey to E and then took a 15-minute recess. Soft drinks and donuts were available at this time. Refreshments were served to reduce fatigue effects, to occupy the unprogrammed time during the recess, to reduce anxiety and

promote an atmosphere of free interaction among Ss prior to the debriefing.

During the recess, E and SLATE author tallied attitude survey and post-test scores and noted those items which indicated a discrepancy for 30% or more of the Ss. These discrepant items became the agenda for the debriefing.

Debriefings were conducted in the carrel room. The SLATE author began each debriefing by reiterating his need for frank, candid, constructive criticism since the author and program were being evaluated, not the Ss. Using the agenda developed during the recess, Ss' interaction was guided towards the problem areas. As specific problems were broached by Ss, E wrote the problems on a poster board large enough to be seen by the group. Debriefings concluded naturally after approximately one hour.

B_1 experimental procedures.--Data collection in B_1 experimental and control treatments were conducted from 7:00 until 10:00 p.m. in the Industrial Arts carrel facility, 115G Erickson Hall, Michigan State University, during Fall term 1970 and Spring term 1971.

B_1 differed procedurally from A_1 and A_2 in several significant ways. First, B_1 used a group presentation mode instead of Ss interacting with SLATE materials on a self-paced, self-instructional basis. The SLATE author operated the AV equipment and Ss were instructed to interrupt the presentation any time they had a question. The ensuing interaction involved the entire group and allowed the SLATE author to establish an immediate consensus on any given problem by asking, "How many of you (Ss) feel that way about X . . .?"

Due to interruptions and SLATE author explanations, total instructional time during prototype (control group tryout) was 98 minutes. This represented a 500% increase over the 17-minute elapsed running time of the prototype self-instructional AV presentation.

Ambient light in the room was a factor in that Ss could not clearly see their workbooks in the dark, nor the screen with the lights on. Since many responses were related to visual discriminations on the slides, the inability to see workbook and screen simultaneously may have adversely affected learning.

The orientation briefing, pre- and post-test scoring, and use of post-test and attitudinal data to develop a debriefing agenda were similar to A_1 and A_2. Moreover, the debriefing itself was procedurally the same. But since much of the information had been discussed earlier in the context of the lesson, B_1 debriefings were typically one-half the length of A_1 and A_2.

Research and Statistical Hypotheses

The following research and statistical hypotheses were tested in all three experimental comparisons; A_1, A_2, and B_1.

H_1: Ss using revised instructional stimuli will show greater mean achievement on post-tests than Ss using prototype (unrevised) instructional stimuli.

H_1: $\bar{X}_e > \bar{X}_c$ $\qquad$ H_0: $\bar{X}_e = \bar{X}_c$

H_2: Ss using revised instructional stimuli will show greater gain score between pre- and post-tests than Ss using prototype (unrevised) instructional stimuli.

H_2: $\bar{X}_e > \bar{X}_c$ $\qquad$ H_0: $\bar{X}_e = \bar{X}_c$

H_3: Percentage of Ss achieving "mastery" (80% correct on post-test) will be greater among Ss using revised instructional stimuli than among Ss using prototype (unrevised) instructional stimuli.

H_3: $\%_e > \%_c$ $\qquad$ H_0: $\%_e = \%_c$

H_4: Ss using revised instructional stimuli will show a greater mean score on measures of attitude regarding effectiveness of instruction than Ss using prototype (unrevised) instructional stimuli.

H_4: $\bar{X}_e > \bar{X}_c$ $\qquad$ H_0: $\bar{X}_e = \bar{X}_c$

Data Analysis and Statistical Treatment

H_1: Involves a comparison between the mean achievement scores on post-test instruments using two independent samples (N=12). Assuming interval data, equal population variances, and normal distribution of the achievement scores, a t test is an appropriate test of significance.

H_2: Involves a comparison between the mean gain score (difference between pre- and post-test scores) using two independent samples (N=12). Assuming interval data, equal population variances and normal distribution of achievement scores, a t test is an appropriate test of significance.

H_3: Involves a comparison of the difference between two proportions; the proportion of Ss achieving "criterion" in the experimental treatment compared to the proportion of Ss achieving "criterion" in the control treatment. The significance of this difference may be computed by determining the standard error of the difference between two uncorrelated proportions, converting this to a z score, and determining the probability of such a z score from the table of the normal curve (Edwards, 1950, p. 77).

H_4: Involves comparison of mean scores between two independent samples (N=12) on measures of Ss' attitude towards the instructional stimuli and total learning environment. Assuming interval data, equal population variance, and normal distribution of the attitudinal scores, a t test is an appropriate test of significance.

Chapter Summary

The descriptive and experimental methodology used to assess the validity, practicability, and efficiency of the MK II model have been described in this chapter. The methodology involved experimenter developed narrative reports of all revision activities with each of three authors.

The experimental design involved three separate field experiments each using a group design. Three prototype SLATE authors were selected to develop revised versions. Ss were volunteers from the SLATE author's course who were stratified into three groups according to Scholastic Aptitude Test (SAT) scores. Four Ss from within each group were randomly assigned to treatments (N=12). Effects of experimental and control treatments were determined by measures of four dependent variables: mean achievement, gain score, percentage achieving criterion, and mean attitude score. Four hypotheses regarding comparison between experimental and control groups were tested at the .05 level of significance. T tests and/or a table of the normal curve were used as appropriate. A schematic of the experimental comparisons is shown in Figure 11.

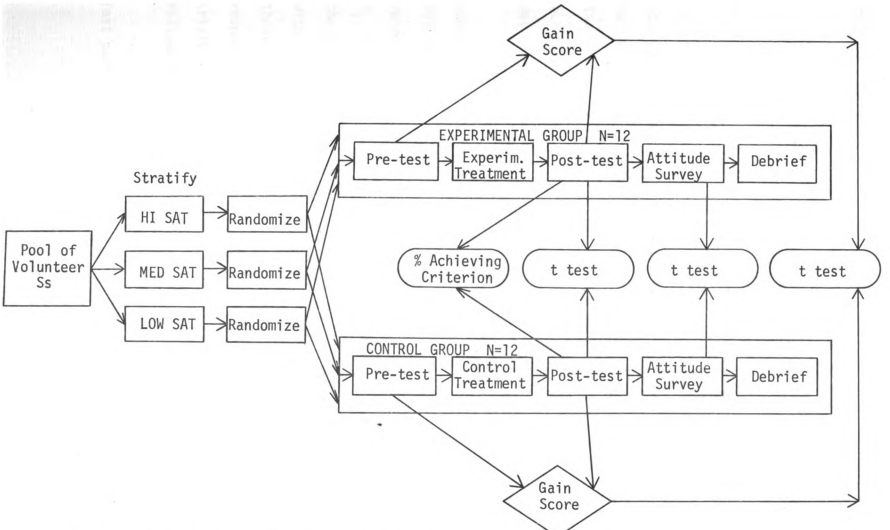


Figure 11.--Schematic of Experimental Comparison Methodology

CHAPTER V

DESCRIPTION AND RESULTS OF FIVE FIELD TRIALS

The purpose of this chapter is to describe the results of five field trials in which the experimenter collaborated with three Michigan State University faculty to apply the MK II model of formative evaluation to revise five prototype SLATEs. Three of these SLATEs were developed by the same author (A) and are designated A_1, A_2, and A_3. The other two SLATEs were each developed by two different authors (B and C) and are designated B_1 and C_1.

The chapter is divided into two sections, descriptive and experimental. The descriptive data are presented first and are organized as follows. Each of the field trials will be described across trials with respect to a given step in the MK II "maxi" model. For example, trials A_1, A_2, A_3, B_1, and C_1 are described with respect to "Logistics." Then the five trials are described with respect to "Data Collection," and then with respect to "Data Analysis" and so forth. This parallel organization should facilitate drawing inferences regarding effects of the model with respect to each of its functions as they were performed during the field trials. The description of the field trial activities in this chapter follows the same sequence as the MK II "maxi" model shown in Figure 8.

The experimental data for each field trial are reported in the last part of this chapter.

Technical Assessment Cycle

The MK II tryout and revision process begins by obtaining feedback from technical consultants. The first part of this chapter describes how the technical assessments were performed in this study. A brief discussion follows presentation of the descriptive data.

Logistics for Consultant Tryouts (Box 1.2)

With respect to selection and briefing of consultants, the experimenter (E) functioned as the instructional media and evaluation instrument consultant in all trials; thus no briefings were required on these topics. Authors A and B used departmental faculty for subject matter consultation and utilized a briefing guide provided by E for this purpose. Author C did not conduct subject matter evaluation since his SLATE had been co-authored; i.e., the content had already been assessed for accuracy and up-to-dateness during development and needed no further evaluation.

Extra copies of prototype materials were not required in any trial since consultants reviewed the materials sequentially. None of the authors had used a storyboard in development of their SLATEs; consequently, no storyboards were available for recording the technical data collected from consultants.

Data Collection on Technical Problems (Box 2.0)

Subject matter assessment (Box 2.1).--Author A conscientiously followed the prescriptions of the MK II model by asking peers to assess the subject matter accuracy, up-to-dateness, treatment and objectives of his three SLATEs. However, it did not appear as though feedback from the subject matter experts resulted in significant revisions to the prototypes. For example, Author A included only minor changes in two scripts based on

feedback from his technical experts. Author B used a graduate student to assess the content of the prototype SLATE. No changes in the lesson resulted from his assessment.

Media assessment (Box 2.2).--In all five trials, E functioned as the media consultant. The prototype SLATEs of each author differed considerably with respect to the media characteristics specified in the model. For example, Author A's SLATEs were of uniformly high technical quality having been produced by a professional staff in the MSU Instructional Media Center. On the other hand, Author B had done much of the photography and all of the sound recording himself using amateur equipment. Consequently, the B_1 prototype was technically inadequate. Many slides, for example, were copied from technical manuals in which the print was so small as to be illegible on the screen. In addition, the audio tape had been recorded with subaudible "beeps" to automatically advance the slides. E recommended eliminating the automatic feature in the prototype since it was likely that students would wish to stop the tape, rewind, and listen to the tape again. Nevertheless, Author B was unwilling (due to lack of resources) to remediate any of the technical problems prior to student tryouts. E did not take issue with this decision since to do so might have jeopardized B's commitment to the project, which was not high to start with.

Author C had some technical discrepancies (such as illegible print on slides) which E felt should be revised prior to student tryout. Nevertheless, Author C was also unwilling to make these technical revisions prior to student use due to lack of resources.

Evaluation instrument assessment (Box 2.3).--Major discrepancies were found in four out of five SLATEs, in that no evaluation instruments of any type had been produced. The authors had not recognized the necessity of evaluating an individual SLATE as an entity and had, instead, relied on mid-term and final examinations to assess students' learning from SLATEs. The first major problem, therefore, was to convince the authors of the desirability of the criterion referenced evaluation procedures stipulated in the MK II model. The rationale was presented and these concepts fairly readily accepted by A and B. Author C, however, remained adamant that pre-post and en route tests were not needed in his SLATEs since he could assess student learning by means of personal observation of their performance on experiments.

With respect to an attitudinal survey instrument, none of the authors had used them before. In light of the time available, all were favorable to using the instrument developed by E.

Problem Analysis and Interpretation (Box 4.0)

The data analysis was simplified since the only remedial action the authors were willing to take initially was to create evaluation instruments, although other technical discrepancies had been noted. The authors were collectively bothered by the high cost of revising the media prior to student tryouts and would not undertake this task. On the other hand, development of evaluation instruments was relatively low-cost, e.g., secretarial help and duplication costs. Even so, Author C was unwilling, despite presentation of the same rationale by E, to adopt the evaluation model stipulated in the MK II model. Purely and simply, the development of pre-post and en route tests from "scratch" is a formidable job, and one

which Author C did not then have time to perform. Furthermore, Author C had not specified his SLATE objectives in behavioral terms so the problem of evaluation instrument development was made even more complex and time consuming.

On the other hand, Authors A and B had fairly specific objectives but no evaluation instruments to assess achievement.

Revision Development (Box 5.0)

Authors A and B, on their own, developed post-tests which were then assessed by E for congruence with SLATE objectives and feasibility within the SLATE environment. The initial instrument on A_1 was far too long (eighty multiple choice and fill-in items) and was reduced by deleting thirty items and using them as part of the en route items. It was suggested by E that feedback be provided students after each major topic in the SLATE, so ten-twenty en route items were embedded in the SLATE workbook. Author B, on the other hand, developed a post-test which did not fully sample the intended objectives, so it had to be revised to include more comprehensive coverage.

After development of the package of instruments for his first SLATE, Author A subsequently became increasingly proficient in producing satisfactory instruments for SLATEs A_2 and A_3. In fact, SLATE A_3 prototype instruments required no technical revision prior to use with students.

Discussion of the Technical Assessment Cycle

The importance of technical assessment prior to student tryouts was highlighted by the deficiencies found in the instructional media and in the evaluation instruments. Unfortunately, none of the three authors were willing to revise technical discrepancies in the instructional media

prior to student tryouts. Their attitude was "let's see if the students complain before we invest any more time or money." Their attitude seemed similar with regard to content since data collected from subject matter experts appeared to have minimal impact. Furthermore, during this phase, consultant assessment of "treatment" (Box 2.1.3) or "objectives" (Box 2.1.4) was non-existent. With a view towards simplification, these steps should probably be deleted.

In sum, the major problems uncovered during the technical assessment were the lack of evaluation instruments. In each case, these discrepancies were attributable to the authors' lack of familiarity with the principles of criterion referenced instruments, gain scores, and en route assessment and feedback techniques.

Student Tryout Cycle

Following technical assessments, the MK II process recycles and begins to focus on student learning problems. This section describes the activities performed within each step of the second cycle. A brief discussion of Cycle 2 follows the descriptive data.

Logistics for Student Tryouts (Box 1.3)

In all trials, the logistics required for student tryouts were initiated while the technical assessment was still in progress.

Select Ss from volunteer pool (Box 1.3.2).--Two authors (A and B) solicited volunteers from their classes. Author A asked for volunteers to participate in a new SLATE a few weeks ahead of the rest of the class, so they could provide feedback at a debriefing and help him improve the lesson. Interestingly, the number of volunteers increased on later trials indicating a favorable student response to this technique.

Author C did not wish to solicit volunteers and go through the randomization and stratification procedure. Instead, he wished to use an intact "quiz section" of ten students which he taught. During the prototype tryout, fourteen students attended the quiz section and all were allowed to participate.

Obtain and set up carrel facility (Box 1.3.4).--A major logistics problem was scheduling tryouts to fit into an already busy carrel facility schedule and scheduling the student tryouts to occur at the optimum time in each ongoing course. Often, SLATEs are hierarchically sequenced, so considerable prerequisite knowledge is required in later SLATEs. Tryouts must, therefore, be scheduled while students are naive with respect to the specified SLATE, yet have achieved the necessary prerequisite concepts. The time frame within which tryouts and revisions had to be complete for A_1, A_2, and A_3 SLATEs was one week due to the hierarchical sequence in this course. The B_1 and C_1 SLATEs had considerably more flexibility with respect to sequencing the tryout in these courses.

Determine data format and tryout procedures (Box 1.3.1).--The way in which the SLATEs were presented to students differed considerably among the five trials, although the debriefing procedures were identical. Author A, for example, had a complete carrel facility capable of handling twelve individual students simultaneously. He elected, therefore, to allow students to interact individually with the prototype SLATE materials during which time he would circulate in the room answering individual questions in a tutorial fashion.

Author B, on the other hand, had no carrel facility whatsoever. Therefore, of necessity, the tryout procedure for his prototype SLATE

used a group presentation mode. That is, the group viewed the slide/tape presentation in a small classroom, but individual students responded in their own workbooks. Individual student's questions were encouraged, but of course the entire group was in "lock step" during the presentation and author response to questions.

Author C had a large carrel facility but did not have the resources to make twelve copies of the prototype SLATE materials prior to student tryout. Therefore, he elected to use the group presentation mode for the tryout session.

In sum, pragmatic considerations such as the availability of carrel facilities and sufficient copies of the prototype SLATE seemed to dictate the presentation procedures used during the student tryouts. There appeared to be little difficulty in obtaining sufficient volunteers to convene a group of twelve Ss stratified by SAT. A major logistic problem, however, was the coordination of tryouts within the carrel facility operating schedule--at the optimum time in the hierarchical sequence of course activities.

Collect Student Tryout Data (Box 3.0)

Conduct group orientation briefing (Box 3.1).--The first activity during all student tryouts was an orientation briefing. The purpose of the group briefing was to inform students of the "ground rules" and to reduce their anxiety so responses would be frank and honest. All three authors conducted satisfactory briefings by following a checklist provided by E for this purpose (Appendix D). Authors A and B made use of student name tags provided by E and encouraged the tryout session to operate on a first name basis. Although name tags were distributed at C_1 tryout,

the room was much larger allowing students to spread out. Author C could not see the students' names, hence did not operate on a first name basis. Moreover, during the initial briefing, it appeared that A and B went to greater lengths to put the students at ease (e.g., told some jokes) and generally established a more relaxed, informal atmosphere.

Collect Individual Tryout Data (Box 3.2)

Achievement measures (Box 3.2.1).--In all trials conducted by authors A and B, criterion referenced, self-scoring pre- and post-tests were administered. As soon as instruments were completed and scored, they were collected by E and scores entered on an item by student matrix or tally sheet. If 30% or more of the students missed an item on the post-test, that question became an agenda item for the debriefing. To discourage cheating, answer keys were distributed as students neared completion of the instruments. In the case of C_1, no achievement data was collected.

Affective/attitudinal measures (Box 3.2.3).--In all trials, the attitudinal instrument developed by E was used. Procedurally, it was administered immediately after students had completed the post-test. After completion, the instrument was collected and scored immediately by E. Transparent overlays of each page were used indicating the "ideal" student response to each item. (The "ideal" response was one at the extreme end of the rating scale, scored as five points.) If a student's response differed by more than two points from the ideal, that item was entered on a tally sheet. If the tally sheet indicated that more than 30% of the students indicated concern on a given item, that item became an agenda item for the group debriefing.

Process and tutorial data collection (Box 3.2.2).--Process data were collected by observing students taking the lesson and answering their questions in a tutorial fashion. Written notes were made by each author as students asked questions and all author/student interactions were tape-recorded by means of a small portable cassette recorder. If more than three students asked a similar type of question, that question became a debriefing topic.

In the case of Author A, all students used the prototype SLATEs in individual study carrels. The author would tutor a student individually when a student signaled a need for help by raising his hand.

In the case of B_1 and C_1, however, the SLATE was presented to the entire group simultaneously, although students responded in individual workbooks. Students were instructed to raise their hand and stop the presentation at any time. During B_1 and C_1, however, students seemed reluctant to interrupt the presentation and often needed an "ice breaker" or student who was willing to ask the first question. In B_1 and C_1, the "ice breaker" function was served by E. That is, after E had observed some nonverbal indicators of confusion (student frowns, looking around, etc.) yet no questions were forthcoming, E took the initiative and asked the author to stop the presentation and inquire if anyone had any questions. Invariably, this stimulated a host of questions and seemed to facilitate subsequent questions from the group.

Another problem in the group mode was room darkening. Often, the SLATE would require a response in the workbook which was impossible to perform with the room lights off. On the other hand, both B_1 and C_1 had a great deal of printed information on slides which was impossible to see with the room lights on. Therefore, it became necessary to switch the room lights off and on, which seemed to make the SLATE presentation very

laborious. Furthermore, when a student asked a question with the room dark, it seemed uncomfortable not to be able to see who was talking. Since the author's response often involved some drawing at the blackboard, gestures, or other visual cues, it was determined to turn on the room lights whenever students asked a question. This convention may possibly have inhibited the frequency of questions during the presentation. That is, students appeared to save up questions; and when a more aggressive student finally asked a question (stopping the presentation) a large number of students would then take advantage of the opportunity and ask questions which had been worrying them.

Although the unit of data collection for process and tutorial data was the individual student, identification of major problems was simple in that a large percentage of students often appeared to need tutorial assistance in similar areas. The direct tutorial interaction provided the lesson author valuable experience in remediating the major difficulties and the tape recording provided a first draft script for the revised version.

Collect Group Debriefing Data (Box 3.3)

Develop agenda (Box 3.3.1).--The purpose of developing an agenda was to form an organized basis for conducting the debriefing. On all five trials, after students finished the presentation, post-test, and attitudinal survey, they were given a 15-minute "break." During this time, E and the lesson author worked at maximum speed to tally individual responses on both the post-test items missed by more than 30% of the group and those attitudinal items in which more than 30% of the group responded more than two points from the maximum. These data were tabulated on the actual

post-test and/or the actual attitude survey instrument. During the break, students were asked not to discuss the SLATEs, but rather to save their comments for the debriefing.

Conduct group debriefing (Box 3.3.2).--This phase of the data collection was critical because here it was assumed that through face-to-face interaction the author would be able to conceptualize the most serious problems in the prototype. Furthermore, it was assumed that the students would suggest solutions to the problems identified. In general, both these assumptions proved valid.

In the case of A_1, A_2, and A_3 SLATEs, the major learning problems were identified prior to the debriefing either on the post-test, attitude survey, or author's notes. However, the debriefing served to explicate these problems in a way which made their solution much more apparent than would have been possible without the debriefing.

For example, students often indicated that they were unclear as to what they were supposed to be learning. These data seem contradictory since the lesson objectives were clearly specified for the student on page 1 of the workbook. After some discussion at the debriefing, however, it developed that students do not usually attend to statements of lesson objectives; e.g., students cited other lessons in which stated objectives were unrealistic. Author A determined, therefore, to include in the revised version a statement of lesson objectives on the tape which would thereby focus student attention to this point.

As another example, students indicated that the post-test was unfair, in that it was not a representative sample of lesson content. This, in spite of the fact that E and Author A had agreed that the

post-test adequately sampled student knowledge with respect to the lesson objectives. After some discussion, it became clear that the problem did not lie in the post-test which did, in fact, test lesson objectives. The problem was in the relative emphasis given certain content in the SLATE--which was not reflected either in the lesson objectives or the post-test. Specifically, 15 minutes of one SLATE was spent on historical development of the cattle industry (with numerous places, dates, and other historical information). Knowledge of historical development was not a major objective of the lesson, consequently only two (out of fifty) post-test items referred to historical development. The students, in the meantime, had been concentrating on memorizing the historical part at the expense of the other concepts. The debriefing, therefore, had explicated the combination of factors which led to this feeling of frustration on the part of students; namely, they didn't read the objectives, and the SLATE content overemphasized that which was not a lesson objective.

In sum, these two examples are illustrative of the value of the debriefing to explicate problems and show their interrelationship. Both these discrepancies (test unfair and didn't know objectives) had been identified by specific items on the attitude instrument. In addition, the post-test had shown a lower than expected performance on critical areas. However, it would have been impossible from these instruments alone to deduce the actual causes and the interacting circumstances. That is, only through the debriefing did it become clear that since the students did not read the objectives and since the SLATE overemphasized historical development, the students concentrated their efforts in this

area and did poorly on other areas of far greater importance. The implications for revision become fairly obvious in light of understanding the problem.

Each debriefing was largely self-directional. As a problem was listed, students would discuss it and propose several alternative solutions. In most cases, shortly after one hour the major problems had been discussed and solutions generated. After this time, students began showing signs of fatigue and the SLATE author clearly recognized areas in need of revision.

Another characteristic of the Author A debriefing sessions was the frankness and honesty of students. Author A made it very clear that grades would not be affected by any remarks made during the debriefing--that there would be no reprisals. Furthermore, if the students were not frank and honest then the whole procedure was a waste of time. Consequently, a very intensive interaction developed in which students often made criticisms which were harsh, hostile, and derogatory. After a while Author A naturally became defensive. In one A_1 session, however, the students sensed that they were being overly critical and abruptly reversed their direction with a series of comments on the positive aspects of the lesson.

In most sessions, there appeared to be a period of time where the students were "sparring" or testing the author to see if he really wanted substantive criticisms. After each new group sensed the author was willing to listen, they often responded with a veritable barrage of discrepancies.

Moreover, when the session had progressed to substantive issues, students would often come forth with "confessions" of how they had "beat

the system" previously and how the author could reduce this contingency. For example, at one session, the students volunteered that they had simply memorized many of the self-scored pre-test answers and drifted through the lesson haphazardly relying on their memory to get them past the 80% criterion on the post-test. They suggested, therefore, that different (but equivalent) items be used on self-scored post-tests to discourage memorization of answers and encourage actual learning. This suggestion was adopted for all subsequent SLATEs.

Interestingly, no matter how vehement the students became during the debriefing, invariably they expressed their appreciation for the opportunity to make an input into their instruction. Furthermore, several students indicated that this was their very first chance to talk face-to-face with the faculty member and this opportunity was appreciated. As the debriefings digressed occasionally, other information was fed into the author with respect to the overall course, as well as the specific SLATE. In several cases, students suggested a different sequence of SLATEs than was presently being used. Their rationale was so logical that it was beyond question. Thus, debriefings can serve a greater need than just the immediate SLATE.

Since Author A handled a total of five debriefing sessions, he became increasingly skilled at this task. Moreover, it was apparent that he was internalizing the feedback since initial prototypes of subsequent SLATEs were far better than his earlier ones. (Higher percentage of Ss achieved criterion on post-tests and higher mean on reactionnaire.) The improved prototype resulted in a lower level of interaction during the debriefings comparing SLATE A_1 with A_2 and A_3. The term "level" is used

here to denote both frequency and intensity of interaction. In short, on later SLATEs the students had less complaints; hence the debriefing interaction was somewhat forced and far less vehement. In fact, on SLATE A_3, the discrepancies noted were not severe enough to warrant any revisions at all.

The quantity of feedback during the debriefing did not seem appreciably different between students of high, medium, or low SAT, but varied primarily with their verbal ability and previous experience with the subject matter. Furthermore, it did not seem that low SAT students were encountering different problems than the high SAT. Possibly during the initial tryout and debriefing sessions the problems identified were of such a gross nature (e.g., not enough space to write in the workbook or wrong color on a slide, etc.) that major problems emerged regardless of the composition of the group.

The fact that the group was composed entirely of volunteers eager to provide feedback was likely to be influential in obtaining the high level of interaction. In addition, the N of twelve seemed ideal to stimulate discussion, yet small enough to allow the less verbal students to participate. Relevant feedback was obtained from each student; yet confirmation was available from others in the group so that the author was aware of a generalized problem.

Author B used essentially similar debriefing procedures as Author A, but the B_1 prototype SLATE had so many serious discrepancies that the student literally destroyed it. This particular SLATE was B's first effort and since professional production techniques were not used, the technical quality of many slides was criticized. Furthermore, as students

asked for replays during the presentation, B had great difficulty in regaining slide/tape synchronization.

Besides technical problems, the major discrepancy in this prototype was quite simply that it did not teach the prerequisite concepts (nomenclature and computation) needed to achieve the terminal objectives; namely, how to read a micrometer. Consequently, on numerous occasions during the presentation the author was required to reteach in front of the group that which had just been shown in the SLATE.

Again, all B_1 major discrepancies were reflected in the attitudinal and post-test instruments. These discrepancies included: irrelevant information, too rapid pacing, not enough practice, poor selection of visuals, and lack of organization. During the debriefing the students in this particular group were very blunt and critical on these points. Although the students contributed a number of highly constructive suggestions or solutions to these problems, the net effect on the author was that his entire prototype was perceived as a complete failure which would have to be redesigned from scratch. Given the amount of time and effort which had gone into the development of the prototype, this type and intensity of feedback clearly discouraged B from engaging in an immediate effort to redesign the unit. In short, the vehemence of the student debriefing engendered by the poor quality of the prototype was enough to discourage the author for several months.

In contrast with the highly interactive and frank sessions enjoyed by A_1, A_2, A_3, and B_1, the C_1 debriefing session seemed strained, guarded and perfunctory. Several reasons for this were hypothesized. First, although C appeared interested in formative evaluation and said the

right words at the introductory briefing, as the presentation continued it became increasingly clear that C did not want criticism, but rather, reinforcement for his prototype efforts. For example, during the audio-visual presentation prior to the debriefing, few students asked questions--in spite of prompting by E. The reason appeared to be that C was slightly irritated by the questions. So rather than providing an explanation, C would often respond with a justification of the presentation or an evasion, such as "that will become clear later."

Following the presentation, data on the student reactionnaire clearly indicated that students were confused; did not know the lesson objectives; did not know what they were responsible for learning from the slides and tape; and had been overloaded with information. When these issues were raised by C, they were introduced by comments such as: "You read the objectives in the workbook and still didn't know what they were?" "At the end of the lesson you still can't tell the major concepts from the supporting facts?" The tone of subtle incredulity in C's voice probably intimidated many students and stopped productive discussion on these problems.

Several other factors may have contributed to the low amount of interaction. The group was primarily first term freshmen who were still in awe of the perceived power of the professor. In spite of words to the contrary, there may have been a fear of reprisals. Furthermore, this try-out session took place on the third day of class before the students had gotten to know C or each other. In addition, the co-author (another senior faculty member) was present in the room; so to speak out was possibly to invite intimidation or reprisals. Also contributing to the

lack of interaction was the larger size of the room (about 50' x 150'). Students preferred to spread out and it was difficult to hear all the comments. A final and possibly a critical factor was a time constraint. The entire presentation and debriefing had to be finished within the 50 minute laboratory period (the presentation itself was 20 minutes long). This tended to encourage clock watching rather than analysis of problems as students appeared restless in the last 15 minutes.

In short, the C_1 tryout session was structured against effective feedback. The combination of a race with the clock, the large room, the novelty of the situation (for freshmen), and the apparent or perceived intransigence of C probably all tended to inhibit productive analysis of the problems identified on the attitude instruments.

In contrast, the debriefing procedures used by A and B were highly effective in explicating the problems identified on the evaluation instruments. Organizationally, in all A and B tryouts the students chose to discuss a specific problem and finish with it before going on to the next one. Consequently, the problem posting procedure was not used. The agenda prepared during the "break" was used to organize and focus the discussion, but not restrict it. In all cases, the authors felt that the total procedure of student tryout followed by a debriefing had produced more than enough data upon which to base a revision. In fact, the problem which occurred was that of information overload. The rapidity of suggestions, skipping to different aspects of the problem, and freewheeling atmosphere tended to obscure the solutions in a barrage of words. For this reason a tape recording was essential for later analysis.

Data Analysis (Box 4.0)

Following each debriefing session, E consulted with each author regarding data interpretation and design of revisions. The first step in each of these consultations was to get the problems out on the table to determine their interrelationships and decide which problems warranted immediate revision.

In the case of C_1, the debriefing had not convinced him of the necessity of any revisions inspite of the data on the attitudinal survey. He maintained that deficiencies in the SLATE would be ameliorated by subsequent lectures, SLATEs and lab sessions in the course. The main problem here was that the student workbook for this SLATE had already been printed and was on sale at the bookstore. Essentially, it was too late to change the objectives, organization, or substantive content of the lesson. C did agree that several examples in the presentatation were clearly at too high a conceptual level for the students, but he contended that exposure to this type of example (a quotation from a professional journal) was a desirable experience in spite of student objection to it.

During this discussion, E became aware that this author's basic position was that faculty, not students, were the final arbiters of what is best for student learning of a subject. Furthermore, faculty designed lectures and labs were seldom, if ever, revised on the basis of student input but rather on input from the professional discipline. Since SLATEs were far more carefully designed than lectures, faculty assume that the prototype SLATEs teach more effectively than any lecture, hence need little or no revision. In any case, any SLATE deficiencies would be remediated via lectures and labs. Since there was no objective assessment of

individual SLATEs, one cannot argue as to their individual effectiveness (or lack of it) except by inferring poor performance on mid-term or final exams. In short, Author C was not interested in pursuing formative evaluation as defined by the MK II model.

On the other hand, the data analysis function performed on the B_1 prototype SLATE revealed so many major problems that the most logical strategy was to abandon the prototype completely and develop a totally new lesson.

It was suggested by E that attention be given to development of realistic objectives followed by a thorough task description and analysis. The resulting flowchart could then be used as a basic outline for the revised version.

The analytic procedures used with Author A were essentially those prescribed in the MK II model. At each session following a student debriefing, all discrepancies were listed in problem posting style. Then these problems were organized into what might be called "strategic" and "tactical" categories. Strategic problems related to major changes in lesson objectives, sequencing, content, or evaluation instruments. "Tactical" problems were of the nature of obtaining better exemplars of specific cattle breeds or clarifying instructions in the workbook. The MK II "maxi" tradeoff analysis procedures were not used formally. Instead, tradeoff and go-no-go decisions were performed by Author A quickly and intuitively. There was simply not enough time to go through the careful analysis specified in the model.

The following list is illustrative of the type of discrepancies which were brought to light at the debriefing and which the author and

E subsequently analyzed to design revisions. These discrepancies were identified in a lesson on cattle breeds:

1. Too much new information too fast.
2. Slides don't exemplify the specific breed being talked about on the tape.
3. Poor examples of specific breeds; e.g., the "Red Poll" was brown and the "Black Angus" was navy blue, a horned breed was shown without horns.
4. Should use simultaneous not sequential presentation of different breeds.
5. Overemphasis on historical development.
6. Critical cues not highlighted on pictures of different breeds.
7. Use more than one shot or example of various breeds.
8. Graph in workbook totally unfamiliar and unusable.
9. Workbook has insufficient space to take notes.
10. If a slide is omitted because there is not a good photo of a breed--tell the students.
11. Have students write own definitions in workbook.
12. Make alternate forms of the pre- and post-test.
13. Do not use black and white pictures of colored breeds.
14. Break the lesson into two parts, foreign and domestic breeds.
15. Exams don't reflect lesson content.

The nature of the student debriefing provided such extensive feedback that Author A usually had a clear understanding of the major problems and some alternative revision hypotheses at the end of the debriefing. The data analysis sessions with E largely involved a review of the feasibility of student generated suggestions with cost and time constraints as critical variables.

In the opinion of the three participating authors, the most odious discrepancies were those involving slide/tape and workbook coordination. Due to this interactive effect, such discrepancies usually received top priority or were disregarded as too complex to undertake under existing time constraints. For example, when students suggested reorganizing one long SLATE which was overloaded with information into two shorter versions, the slide/tape/workbook interaction prevented doing this easily so the idea had to be abandoned.

Design of Revisions (Box 5.0)

The most extensive revision was carried out by Author B who completely redesigned his SLATE. The major organizing heuristic for this redesign was provided by a student who suggested teaching the concept of 1/1000's of an inch by developing the analogy between 10 dollars and 1000 pennies. At E's suggestion, specific behavioral objectives were developed, and task description and analysis performed. In short, with Author B, virtually all the revision variables stipulated in step 5.0 were involved; i.e., objectives, evaluation instrument redesign, inclusion of student response and feedback, and a complete reorganization and treatment of the content.

Author A, on the other hand, obtained specific redesign information from students and essentially knew where and how the unit would be revised immediately after the debriefing. The cafeteria function of step 5.0, therefore, was somewhat negated in that the menu had already been largely planned. E provided very little guidance regarding message design due to the complexity of the information. Organization and sequence ideas had already been provided by students.

A major area in which students did not seem able to consistently provide redesign information was in evaluation instruments. Students could quickly point out ambiguities, inconsistencies, and unfairness of examinations. However, they did not seem as able to provide viable solutions to these type of problems as they could in other areas. Therefore, one of the major redesign efforts following a debriefing was invariably the design of items to replace those which were causing students problems, or how to revise items to better reflect the content of the lesson.

In sum, regardless of the formalized model, revisions were designed intuitively based largely on the adaptations of student-provided ideas.

Recycle (Box 6.0)

In each case, integration of revisions within the original version was a major problem particularly if multiple copies of the SLATE were needed. As much time was spent organizing the revisions into existing SLATEs as was spent on the analysis, design, and production of the revisions. For example, replacing pages from bound copies of workbooks or inserting new slides or changing the order of old slides proved to be extremely time consuming. This type of "busy work" became very frustrating to authors, yet often they are the only person who knows the lesson well enough to integrate the revisions.

The decision to recycle (tryout the revised version again) seems to be made purely on pragmatic rather than pedagogical grounds. For example, suppose the SLATE performance criterion was established at 80/80, i.e., 80% of the students achieve 80% on the post-test. If after seeing data to the effect that the 80/80 criterion had not been achieved, the

author may very likely rationalize that the first revision was "good enough" on the grounds that it is not practical to spend too much time on one SLATE when there are many others which are considerably worse. This argument was presented by Author A on SLATE A_3 and it was difficult to argue against this position. Essentially the commitment to achieving an 80/80 or 90/90 performance criterion may gradually attenuate, so prototype SLATE operationally ready criteria may become very flexible depending on author workload, resources, and other situational factors.

Discussion of Data From Student Tryout Cycle

The major difference between the several field trials was the mode in which the instructional stimuli was presented. In general, for purposes of presenting the prototype SLATE and obtaining feedback on learning problems, the group mode seemed much more awkward and cumbersome than the individual mode. The inherent "lock step" of the group was probably frustrating to many students who did not appreciate listening to other people's questions. Moreover, the author was thwarted from taking good notes while operating the projection and tape recording equipment. Also, data relative to SLATE instructions and students' ability to operate the carrel equipment were not available, and turning lights on and off was distracting.

Another disadvantage of the group presentation mode was that the group often built up momentum as greater numbers of students concurred on a given problem. This phenomenon occurred in both B_1 and C_1. Then as more problems emerged, the authors attempted to defend the presentation. The defensive posture quickly intimidated students from raising additional substantive questions. In other words, when authors

encouraged a freewheeling interaction, the student feedback seemed increasingly derogatory. Authors then unconsciously defended their presentation instead of focusing on the problems.

On the other hand the group mode did offer some advantages. First of all, the group presentation was inexpensive. Furthermore, during the presentation the author could observe the entire group for nonverbal cues of boredom or confusion and could stop the presentation to determine the problem. Moreover, when a question was asked by a student, the author could quickly ascertain whether this problem was unique to the student who asked the question or whether the problem was shared by the group. Thus, the author could obtain consensus simply by asking, "How many of you had the same problem?" If a large number of students concurred, the author immediately knew that a revision was warranted. As mentioned earlier, such corroborative feedback from several students increased the liklihood that the author would revise the particular segment.

This concludes the description of the process by which the experimental treatments in this study were produced. The next part of this chapter reports the findings from the three field experiments in which prototypes were compared to original SLATEs to assess the effectiveness, validity, and feasibility of the MK II model.

Experimental Data From Field Trials

Four hypotheses were investigated in each of three field experiments (A_1, A_2, and B_1). Findings are organized so that immediately following each hypothesis, data relevant only to that hypothesis from each field trial are presented.

HYPOTHESIS 1: STUDENTS USING REVISED INSTRUCTIONAL STIMULI WILL SHOW GREATER MEAN ACHIEVEMENT ON POST-TESTS THAN STUDENTS USING PROTOTYPE (UNREVISED) INSTRUCTIONAL STIMULI.

Data relevant to this hypothesis are presented in Table 7. In the case of SLATE A_1, the calculated t ratio of 2.842 was greater than 2.508, the tabled value of t at the .01 level of significance, 22 df, using a one tailed test. Since the calculated statistic exceeds the tabled value, the null hypothesis was rejected and alternative hypothesis 1 accepted.

In the case of SLATE A_2, the calculated t ratio of 2.071 was greater than 1.729, the tabled value at the .05 level of significance, 19 df, using a one tailed test. Again, since the calculated statistic exceeds the tabled value, the null hypothesis was rejected and alternative hypothesis 1 accepted for this field trial also.

In the case of SLATE B_1, the calculated t ratio of 3.796 exceeds the tabled value of 2.650, which occurs at the .01 level of significance, 13 df, using a one tailed test. Since in this case the calculated statistic is again greater than the tabled value, the null was rejected and the alternative hypothesis accepted.

Discussion of Findings Relative to Post-Test Achievement

These data clearly show marked improvement in student achievement on post-tests in all three field experiments. This result was, of course, precisely the reason for the tryout and revision efforts. In spite of the small N in each treatment, the results are unequivocal. It should be noted that only those test items common to both experimental and control group post-tests were used to make the necessary calculations.

Table 7.--Comparison of Experimental and Control Treatment Post-Test Scores

	Treatment	N	Max. Points	$\bar{X}$	σ	DF	Tabled T	Calc. T	Significance
SLATE A_1	Control	12	47	37.17	3.87	22	2.508	2.842	$P<.01$
	Experimental	12	47	42.33	4.62				
SLATE A_2	Control	12	40	29.33	5.58	19	1.729	2.071	$P<.05>.01$
	Experimental	9	40	33.44	6.60				
SLATE B_1	Control	7	15	9.86	2.92	13	2.650	3.796	$P<.01$
	Experimental	8	15	14.25	.83				
SLATE A_3	Control	12	46	43.22	4.71				
	Experimental	NA →							
SLATE C_1	Control	14	NA →						
	Experimental	NA →							

NOTE: These data based on scores from test items common to both prototype and revised instruments.

This degree of improvement between prototype and revised versions may be partially attributed to the fact that somewhat less information was presented in revised versions. Furthermore, information which was critical in terms of facilitating student mastery (80% on post-test) was emphasized by redundancy, voice inflection, and embedded criterion test items (equivalent, not identical) on the revised versions. Essentially the presentation was sharpened, focused, and delimited to facilitate the desired learning outcomes.

HYPOTHESIS 2: STUDENTS USING REVISED INSTRUCTIONAL STIMULI WILL SHOW GREATER GAIN SCORE BETWEEN PRE- AND POST-TESTS THAN STUDENTS USING PROTOTYPE (UNREVISED) INSTRUCTIONAL STIMULI.

Data relevant to this hypothesis are presented in Table 8. In the case of SLATE A_1, the calculated t ratio of 2.711 exceeds the tabled value of 2.508 which occurs at the .01 level of significance, 22 df, when using a one tailed test. Since the calculated t ratio exceeds the tabled value, the null hypothesis was rejected and hypothesis 2 accepted.

In the case of SLATE A_2, the calculated t ratio of 1.024 did not exceed the tabled value of 1.729, which represents the more commonly used .05 level of significance (19 df, one tailed test). Since the calculated value was smaller than the tabled value, the null hypothesis cannot be rejected and no significant difference between the mean gain scores has been established in this trial.

In the case of SLATE B_1, the calculated t ratio of 2.701 was greater than the tabled value of 2.650, which occurs at the .01 level of significance, 13 df, when using a one tailed test. Since the calculated value exceeded the tabled value, the null was rejected and hypothesis 2 above accepted for this trial.

Table 8.--Comparison of Experimental and Control Treatment Gain Scores

	Treatment	N	Max. Points	$\bar{X}$ Pre	$\bar{X}$ Post	$\bar{X}$ Gain	DF	Tabled T	Calc. T	Signif.
SLATE A_1	Control	12	47	23.33	37.17	13.83	22	2.508	2.711	P< .01
	Experimental	12	47	21.25	42.33	21.08				
SLATE A_2	Control	12	40	14.75	29.33	14.58	19	1.729	1.024	No
	Experimental	9	40	16.44	33.44	17.00				
SLATE B_1	Control	7	15	3.72	9.86	6.14	13	2.650	2.701	P< .01
	Experimental	8	15	5.00	14.25	9.25				
SLATE A_3	Control	12	50	26.33	43.22	16.89				
	Experimental	NA→				→				
SLATE C_1	Control	14	NA→			→				
	Experimental	NA→				→				

NOTE: These data based on scores from items common to both prototype and revised instruments.

Discussion of Findings Relative to Mean Gain Score Data

With two of the three trials resulting in significant differences in gain scores, there remains strong evidence that the model and attendant procedures were capable of identification and remediation of student learning problems.

In the case of A_2 where no significant difference occurred, feedback during the debriefing identified a major reason why students did not do better, particularly with regard to gain scores. What occurred was that in the previous tryout of A_2, students encountered difficulty making visual discriminations between fat content of cattle carcasses. The problem was pinpointed as overexposed slides which washed out critical color cues. For the revised version these particular slides were reshot at the proper exposure and changed on the post-test. Due to an oversight, insufficient copies were made so that only the post-test had the properly exposed slides. When the students took the post-test, the properly exposed slides were reactive in that several students were misled into thinking the particular carcasses shown were not fat due to the deep red color. Many of these students had guessed correctly on the pre-test due to the washed out color. The combination of a lucky guess on the pre-test and being fooled by the deep color on the post-test was sufficient to attentuate gain scores. During this debriefing Author A became aware of the fact that many students were attending solely to color, rather than other equally relevant cues. He, therefore, decided to include additional instruction on the tape in a subsequent revision.

The important point to be made here, however, is that while the gain scores showed no significant differences in one case, the MK II

procedures were still able to pinpoint the problem so remedial action could be taken.

HYPOTHESIS 3: THE PERCENTAGE OF STUDENTS ACHIEVING "CRITERION" (80% CORRECT ON THE POST-TEST) WILL BE GREATER AMONG STUDENTS USING REVISED INSTRUCTIONAL STIMULI THAN AMONG STUDENTS USING PROTOTYPE (UNREVISED) INSTRUCTIONAL STIMULI.

Data relative to this hypothesis are presented in Table 9. The test of this hypothesis involved comparing the percentage of students achieving criterion in the experimental group to the percentage achieving criterion in the control group. A z score value was calculated for the difference between the percentages and the probability of this z score determined from the table of the normal curve.

In the case of SLATE A_1, a z score of 1.879 was calculated based on a 33.27% improvement in Ss achieving criterion. Reference to the table of the normal curve indicates the probability of a z as large or larger than 1.879 to be .0294, one tailed. In terms of a significance level of .05, a z of 1.879 must be considered significant. Therefore, the null hypothesis was rejected and hypothesis number 3 accepted.

In the case of A_2, only 66.6% of the students in the experimental group achieved the criterion performance level on the post-test. This figure represents only a 8.27% improvement over the control (unrevised) version. The calculated z was .3857, which does not nearly approach significance. Therefore, in A_2 the null hypothesis cannot be rejected and no significant difference has been established between the percentage of A_2 students achieving criterion in experimental and control groups.

The situation was radically reversed in the case of B_1, however. Here, 100% of students in the experimental group achieved the criterion

Table 9.--Comparison of the Proportion of Students Achieving 80% Criterion on Post-tests Between Experimental and Control Treatments

		No. Ss Failing to Meet Criterion	No. Ss Achieving Criterion	Total Ss	% Ss Achieving Criterion	% Difference Between Experimental and Control	Standard Error	Z Score	Significant
SLATE A_1	Control	5	7	12	58.33%	33.27%	.1770	1.8790	Yes P<.05 (P=.0294) 1 Tail
	Experimental	1	11	12	91.60%				
SLATE A_2	Control	5	7	12	58.33%	8.27%	.2144	.3857	No
	Experimental	3	6	9	66.60%				
SLATE B_1	Control	4	3	7	42.85%	51.15%	.5715	2.4960	Yes P<.01 (P=.0064) 1 Tail
	Experimental	0	8	8	100.00%				
SLATE A_3	Control	2	7	9	77.70%	NA →	→	→	→
SLATE C_1	Control	NA →	→	→	→	→	→	→	→

performance, whereas only 42.85% did so during the control group tryouts. The resulting difference is 57.15%, which calculates to be a z score of 2.496. The table of the normal curve indicates the probability of a z of 2.496 or larger to be .0064. This z score is therefore significant beyond the .01 level allowing rejection of the null and acceptance of hypothesis 3.

Discussion of Findings Relative to Percentage of Students Achieving Criterion

In two cases, A_1 and B_1, a large percentage of students achieved the 80% criterion during the experimental treatment. This reflects remediation of both organizational and content emphasis problems as well as elimination of poor evaluation items. The improved student performance in B_1 was remarkable in that 100% achieved criterion in 47 minutes instructional time, as opposed to 42.85% at criterion after one and one-half hours instruction during the prototype. (This SLATE had been completely reorganized to closely follow suggestions given by students at the prototype debriefing.)

The exceptional case again was SLATE A_2 which only showed 8.27% improvement in percentage of students achieving the 80% criterion. Part of this relatively poor showing could be attributed to confusion on the post-test items related to discrimination between types of cattle carcasses. Again, the use of "properly" exposed slides misled students into selecting the wrong answers based on color alone. Another problem with this SLATE was transfer of training combined with satiation. Students were expected to learn a number of complex anatomical discriminations based primarily on line drawings in their workbook. Yet they were tested

on these concepts using actual photographs of carcasses. Since they had been given insufficient practice in making these discriminations on photographs, many were unable to perform this task satisfactorily on the post-test. Furthermore, there was a satiation or fatigue factor operating. Many students complained that they had seen so many beef carcasses in the SLATE that they all began to look alike; hence on the post-test they just "gave up."

Again, the interesting phenomenon regarding SLATE A_2 was that the MK II procedures successfully provided insight into why the data showed no significant difference.

Included in Table 9 is the percentage of students achieving criterion for SLATE A_3. It can be seen that 77.7% did achieve criterion when using the prototype; hence the author felt justified in not making any further revision.

HYPOTHESIS 4: STUDENTS USING REVISED INSTRUCTIONAL STIMULI WILL SHOW GREATER MEAN SCORE ON MEASURES OF ATTITUDE REGARDING EFFECTIVENESS OF INSTRUCTION THAN STUDENTS USING PROTOTYPE (UNREVISED) INSTRUCTIONAL STIMULI.

Data relative to this hypothesis is presented in Table 10. In the case of SLATE A_1, the calculated t ratio of 2.539 was greater than the tabled value of 2.508 occurring at the .01 level of significance, 22 df, when using a one tailed test. Since the calculated t ratio exceeded the tabled value, the null hypothesis was rejected and hypothesis 4 accepted.

In the case of SLATE A_2, the calculated t ratio of .496 did not even approach the tabled value of 2.539 found at the .05 level of significance, 19 df, when using a one tailed test. Since the calculated value of t was smaller than the tabled value, the null hypothesis was

Table 10.--Comparison of Experimental and Control Treatment Mean Attitudinal Scores

	Treatment	N	Max. Points	$\bar{X}$	σ	DF	Tabled T Ratio	Calc. T Ratio	Signif.
SLATE A_1	Control	12	135	95.17	12.20	22	2.508	2.539	$P<.01$
	Experimental	12	135	106.58	8.28				
SLATE A_2	Control	12	135	105.00	9.82	19	2.539	.496	No
	Experimental	9	135	107.22	9.29				
SLATE B_1	Control	7	135	88.86	8.72	14	2.650	4.101	$P<.01$
	Experimental	8	135	112.00	11.26				
SLATE A_3	Control	12	135	103.44	10.31				
	Experimental	NA →			→				
SLATE C_1	Control	14	135	95.64	9.64				
	Experimental	NA →		→					

NOTE: 27 item 1-5 rating scale used.

retained and no significant difference was established between the attitudinal instruments in these treatments.

With respect to SLATE B_1, the calculated t ratio of 4.101 was far greater than 2.650, the tabled value occurring at the .01 level of significance, 13 df, when using a one tailed test. Since the calculated value of t was greater than the tabled value, the null hypothesis was rejected and hypothesis 4 accepted.

Discussion of Findings Relative to Attitudinal Survey Instrument Data

Again, two of the three SLATEs resulted in significant differences in the mean scores on post instruction attitudinal survey instruments. Of particular note was SLATE B_1, which showed the greatest change in scores for all trials ($\bar{X}_C$=88.85; $\bar{X}_E$=112.0). The relatively low initial score could be attributed to a number of factors, primarily lack of preparation, organization, and technical problems which caused undue student frustration. Needless to say, the revised version was precisely organized and thoroughly reviewed to avoid technical problems.

The deviant again was SLATE A_2, which showed very little difference in student attitudes between experimental and control versions. Note, however, that the initial rating of the prototype was unusually high (105.00). In fact, this rating approached the rating achieved by Author A in the revised version of an earlier SLATE (A_1=106.58). A mean score of 105.00 could easily be interpreted to mean that students were generally pleased with the presentation.

While the overall attitudinal rating of 105.00 was unusually high for a prototype, student achievement on this SLATE was unspectacular (66% achieved criterion). The revision hypothesis one might have drawn from

these data was that the SLATE instruction per se was satisfactory, but the pre- and post-tests needed revision. This hypothesis was corroborated by students during the prototype debriefing.

Data from all trials indicated that when the particular attitude survey used in this study showed a mean score above 100.00, the SLATE was approaching operational readiness. This heuristic was based on observations of eight tryout and debriefing sessions where this instrument was administered. Typically, when the instrument scores were over 100.00, the debriefings were not nearly as interactive or critical of the lesson as when scores were lower.

Summary of Findings

The field test portion of this study investigated four hypotheses in three field experiments. In two cases (A_1 and B_1) comparisons between revised and original versions produced significant differences ($P < .01$) on all dependent variables including: (1) achievement on post-test, (2) gain score, (3) percentage achieving criterion, and (4) post instruction attitudinal assessment. These differences were all significant at the .01 level; hence all four hypotheses were accepted for SLATEs A_1 and B_1.

The case of A_2 was considerably different. Here, a significant difference ($P < .05$) was found only on one dependent variable--post-test achievement. The revised version produced no significant differences on measures of gain score, attitudinal assessment, or percentage achieving criterion. These data are not interpreted to indicate a basic flaw in the MK II model. On the contrary, such data simply corroborate the need for further revisions as suggested during the A_2 debriefing. These findings are summarized in Table 11.

Table 11.--Summary of Findings

	Dependent Measures			
	Post-Test	Gain Score	Percentage Achieving 80% Criterion	Student Attitudes
SLATE A_1	$P < .01$	$P < .01$	$P < .05$	$P < .01$
SLATE A_2	$P < .05$	NSD	NSD	NSD
SLATE B_1	$P < .01$	$P < .01$	$P < .01$	$P < .01$

CHAPTER VI

SUMMARY AND CONCLUSIONS

Overview

This concluding chapter has four sections: (1) a summary of the development and validation of the MK II model, (2) major conclusions of the study, (3) heuristics related to use of the model, and (4) recommendations for further research. Heuristics are included in the present study because: (1) the very small N upon which these generalizations are based preclude more definitive statements, and (2) these heuristics will facilitate use of the model by instructional developers.

Summary of the Development and Validation of the MK II Model

The purpose of the present study was to explicate the formative evaluation component of the instructional system development process. This explication included development and validation of a flowchart model and exploration of means by which systematic feedback from students could be included as an integral part of development of new instructional components.

The MK II flowchart model provided a framework for formative evaluation consisting essentially of student participation in a common instructional experience followed by a debriefing at which participants identified, discussed, and proposed solutions to instructional deficiencies in the prototype lesson. The first, or MK I model, was derived from a review of the literature and revised on the basis of interviews with seven

faculty members. The resulting MK II model was validated in five field trials involving three Michigan State University faculty and five prototype SLATEs.

These field trials had both descriptive and experimental components representing the two types of research objectives of the study. The first type of objective focused on describing and understanding the dynamics of the process through which the experimental (revised) versions were developed. In this study, the experimental versions evolved on the basis of activities prescribed in the MK II model, particularly the feedback from student groups using prototype SLATEs. The second research objective related to comparing experimentally student achievement and attitudes between revised and unrevised versions so that general statements could be made regarding the validity, feasibility, and effectiveness of the model.

While in five trials the authors started using the model, in only three trials did the authors follow through far enough to develop revised versions of prototype SLATEs. Thus only three experimental comparisons between prototype and revised versions were completed, although all five trials were described as far as they went and were used as a basis for conclusions and recommendations.

Conclusions

Conclusion 1: USE OF THE MK II MODEL LED TO DEVELOPMENT OF REVISED SLATES WHICH WERE MORE EFFECTIVE THAN PROTOTYPE VERSIONS.

In three separate field experiments, the data clearly showed statistically significant differences favoring the revised versions on four dependent variables: mean achievement, gain score, percentage of students achieving criterion, and student attitudes. Since the MK II model

prescribed the pattern of activities leading to identification and revision of deficiencies in prototype SLATEs, and since the data strongly favored the revised versions, it is reasonable to infer that under conditions similar to those in the three field trials that the MK II model could be an effective tool in identification and remediation of major instructional problems in prototype SLATEs.

Conclusion 2: THE STUDENT GROUPS ORGANIZED WITHIN THE FRAMEWORK OF THE MK II MODEL WERE ABLE TO: (1) IDENTIFY MAJOR DEFICIENCIES IN PROTOTYPE SLATES, AND (2) SUGGEST EFFECTIVE REMEDIATION HYPOTHESES FOR MOST SUCH DEFICIENCIES.

In three field experiments where prototypes were revised, the student group debriefing technique stipulated in the MK II model was highly effective in facilitating identification of major learning problems in the prototypes. Furthermore, the major revisions suggested by the group were incorporated into the revised versions. Since revised versions were superior to the prototypes on nine out of twelve measures, it may be inferred that student groups as constituted in the MK II model were effective in facilitating identification of major problems and in suggesting appropriate revisions. However, student groups seemed less able to provide effective revision hypotheses when deficiencies occurred in prototype instruments.

Conclusion 3: IT IS DIFFICULT TO PREDICT A PRIORI THOSE SITUATIONS IN WHICH THE MK II MODEL WILL PROVE EFFECTIVE.

Because of the interaction of social and psychological variables over which the model has no control, the overall effectiveness of MK II procedures will vary from situation to situation. For example, in the present study three authors agreed to use the MK II model to revise their prototype SLATEs. In actuality, only two authors did so. The precise

reasons for this are unknown, but, in effect, there are so many confounding variables operating in any given formative evaluation situation that the best one can expect from the use of MK II procedures is to increase the probability of remediating major discrepancies in a prototype unit of instruction. On the other hand, the present study provided some insights into several clusters of these confounding variables which would be amenable to further research.

One set of variables having an enormous effect on data collection and problem solving are the group debriefing dynamics; e.g., the size and distribution of abilities within the group, the interpersonal communication skills of both the students and SLATE author, the perceived objectives of the group, unresolved feelings of fear and distrust, and the structure, organization, and/or ground rules of the group.

In the present study, another dynamic factor was observed to operate, namely, the perceived quality of the SLATE learning experience (either prototype or revised). The SLATE learning experience functions essentially as a common experiential referent for both students and author. If the experience was unhappy, frustrating, and/or boring for the students, they rapidly became hostile, derogatory, and vehement in their comments. Furthermore, the groups appeared to develop a "momentum" phenomenon. If they got started on a derogatory theme, they kept going and the comments became increasingly derogatory until the author was forced to become defensive and terminate the discussion.

Yet another set of variables interacting with the group dynamics was the personality and motivation of the author; specifically, how committed was he to the principle of tryout and revision, and how much

criticism was he willing to endure in pursuit of this principle? In the case of Author A, he was able, repeatedly, to handle a number of derogatory comments and still not become defensive enough to impede the debriefing or to abandon the whole idea of revision based on student feedback. In the case of Author B, however, the prototype SLATE was so ineffective and the derogatory comments of students so devastating that by the end of the debriefing the author was simply unwilling to continue the process for the seemingly ungrateful students. Several months elapsed before the author was willing to continue the developmental work. In the case of Author C, he appeared unwilling or unable to handle many derogatory comments and very often closed off discussion prematurely.

In short, a number of relatively unpredictable factors could quickly negate the most carefully developed feedback and revision procedure. Until the dynamics of the process are more clearly understood, it would be difficult to specify the conditions under which the MK II model would be as effective as it was in the present study.

Conclusion 4: OBTAINING FEEDBACK THROUGH MK II PROCEDURES MAY SERVE AN INSTRUCTIONAL DESIGN TRAINING FUNCTION WHICH MAY RESULT IN IMPROVED QUALITY OF SUBSEQUENT PROTOTYPES.

In this study, Author A developed three prototype SLATEs and revised two of them. The third SLATE was not revised because on the first tryout students met the established 80% level of performance on post-tests and showed no major attitudinal problems on the attitudinal survey instrument ($\bar{X}$=105.0). In contrast, Author A's first prototype SLATE was the least effective. It had the lowest percentage achieving criterion, the lowest gain score, the lowest attitudinal rating, and the most vehement student debriefing. The second prototype SLATE fell in between the first

and third with respect to scores on measures of learning and attitude and attitude intensity of student debriefing. Since these SLATEs were developed sequentially within a two and one-half month time period, it was possible that a major variable influencing subsequent SLATE design was student feedback obtained through use of the MK II procedures.

It appeared that in developing SLATEs A_1, A_2, and A_3, Author A learned not to make the same mistakes twice. For example, when students criticized poor exemplars, misemphasis of content, lack of practice in making discriminations, or use of line drawings where a color photograph was needed, Author A seemed able to remember these criticisms and not make similar mistakes on subsequent designs.

It should be pointed out that previous to this study, Author A had designed ten SLATEs (see Appendix F) which were currently used in his course. These ten SLATEs were largely in prototype configuration since Author A had not previously obtained systematic feedback from students regarding instructional problems. It seemed reasonable to assume that prototype SLATE A_1, his first SLATE in this study, was similar in quality to his ten previous SLATEs. If this assumption was valid, it seemed fair to infer that some of the marked improvement in his design ability on A_2 and A_3 was a result of internalizing principles obtained through formative evaluation feedback. While this conclusion may be questioned, having as its basis an N of one, it nevertheless is supported by data from this study.

Conclusion 5: REVISING SEVERAL INSTRUCTIONAL UNITS USING MK II PROCEDURES MAY PROVIDE SUFFICIENT TRAINING TO ENABLE FACULTY TO CARRY OUT FORMATIVE EVALUATION LARGELY INDEPENDENT OF INSTRUCTIONAL DEVELOPMENT CONSULTANTS.

An adjunct objective of the present study was to ascertain whether faculty familiar with MK II procedures could carry out a successful formative evaluation independent of E or other instructional development consultants. After revising two SLATEs, Author A was asked to conduct the formative evaluation of his third SLATE independent of E or other instructional development consultants. Author A agreed, stating he felt confident regarding data collection, e.g., designing instruments and conducting the debriefings. However, he was somewhat unsure of the data interpretation and revision design aspects. The agreement was that Author A would carry out as much of the formative evaluation as possible by himself and contact E when consultation was needed.

Author A did, in fact, conduct the data collection phase entirely independent of E. E observed the debriefing but did not interact with A or any students. After the debriefing, E briefly discussed the data with A, but it was clear that the discrepancies identified were of a minor nature and did not warrant revision. Based on evidence from this study, it was concluded that revising two to three lessons using MK II procedures may provide sufficient training to enable faculty to carry out subsequent formative evaluation largely independent of instructional development consultants.

Conclusion 6: MK II PROCEDURES MAY PROMOTE A SERENDIPITY EFFECT IN WHICH SPONTANEOUS FEEDBACK FROM STUDENTS MAY LEAD TO: (1) REVISION OF A LARGER INSTRUCTIONAL SYSTEM, AND (2) FACILITATE STUDENT-FACULTY INTERPERSONAL RELATIONSHIPS.

While no formal attempt was made to gather data relative to curricular goals, perceived value of course content, or sequencing of course content, in two field trials these types of data spontaneously emerged during the debriefings. In these trials, students continually questioned the relevancy of the content and suggested changes in sequence. This unsolicited feedback, having been strongly reiterated in consecutive debriefings, suddenly triggered in Author A the realization that the students were right--that the course and curricular goals were largely irrelevant to these students' professional and intellectual needs. The fact of the matter was that students were being taught many concepts simply to please faculty colleagues. Author A subsequently revised his course objectives and sequence and now advocated revision of the departmental curriculum. Thus through a series of fortuitious events, a much larger instructional system than the SLATE was revised. Moreover, it was clear from comments offered by many students as well as authors A and B that the group debriefing was an excellent vehicle to become personally acquainted and promote much improved student-faculty relationships.

Conclusion 7: THE GROUP DEBRIEFING (FACE-TO-FACE INTERACTION) PROVIDES POWERFUL, NATURALISTIC DATA ON DISCREPANCIES BUT MAY HAVE A TRAUMATIC EFFECT ON SOME AUTHORS.

Prior to the tryout of their prototype SLATE, each of the three authors participating in this study were skeptical as to whether the group tryout procedure would be valuable. They doubted whether the nature of the information obtained would be worth their investment of time. At the conclusion of the first debriefing, each author indicated that there was no question that the nature of the information was extremely valuable in terms of revising the prototype. However, the experience had been somewhat

traumatic. For example, when a student told an author face-to-face such things as: "The lesson objectives were not clear" or "The lesson content emphasized one thing while the exam emphasized another"--the authors found this feedback uncomfortable but honest. The student who raised the point was probably sincere. He was telling his reaction to the unit. Then, as additional students corroborated the point being made, the cumulative effect began to make an enormous impact on the author. One might say, the author began to "really believe" after a number of students told him the same thing. Stated another way, it was impossible to deny this information. One could not argue with the students or somehow ignore the discrepancies which were discussed. These discrepancies were very real to the students and they became, through the interaction, very real to the author.

As the discrepancies gradually unfolded, the author began to recognize the magnitude of his errors and a sense of frustration emerged. As the students proposed solutions to these discrepancies, the author (who must do the work to change the unit) saw himself rapidly becoming inundated with more work, whereas, he thought he was through.

The net result was that a great deal of valuable data was produced by means of a somewhat traumatic experience. The degree of trauma varied between authors and was inferred to some degree by their post-debriefing behavior. For example, Author A began revisions of each prototype the next day; Author B postponed further development approximately three months; and Author C would not consider revisions at all. Moreover, none of the authors wished to listen to the tape recording of the debriefing during data analysis with E.

This experimenter hypothesized that the degree of trauma was a function of: (1) the author's tolerance to criticism; (2) how ineffective the prototype was, i.e., how critical were the students; and (3) the author's previous experience with leading problem solving groups. Author A, for example, who conducted a total of five debriefings indicated that he was "less shook up" at the later debriefings than he was at the first one. At the later debriefings, his basic prototype SLATEs were better; he had become somewhat desensitized to student criticism and had gained experience in conducting the group. Author B also indicated he was far less "bothered" at the second debriefing.

In short, the nature of the group debriefing interaction was intense and frank. Authors using this technique for the first time are likely to find the data extremely valuable, but may find the overall experience traumatic. As additional experience in handling the group is obtained, however, the traumatic element seems to diminish as desensitization takes place. Such desensitization was reported by Author A, who conducted a total of five debriefings (two debriefings apiece in A_1 and A_2; one debriefing in A_3).

Conclusion 8: SEVERAL MODELS OF FORMATIVE EVALUATION VARYING IN DEGREE OF DETAIL ARE REQUIRED TO CARRY OUT THE PROCESS. A SIMPLIFIED VERSION IS NEEDED FOR TRAINING AND ORIENTATION PURPOSES BUT A DETAILED VERSION MUST BE AVAILABLE TO EXPLICATE SPECIFIC PROCEDURES.

In the present study, the principles of formative evaluation were new to the three participating authors. Consequently, to obtain author commitment to these principles, E conducted an individualized training program using the seven step "mini" MK II model. This model was used because feedback during development of the MK I version indicated that faculty

became apprehensive and had difficulty conceptualizing the process when confronted with a complex flowchart model. The "mini" model oversimplified the process yet highlighted the two alternative strategies of obtaining feedback from experts and students.

During this study, participating authors were never shown the "maxi" version, but instead E provided the necessary information verbally. In effect, E had memorized the "maxi" version and was able to fill in the detail in each step of the "mini" model as required.

In general, data collected from interviews early in the present study and through interactions with the three participating authors, tended to show that the principle of formative evaluation was not widely used by authors during development of new instructional units. Therefore, to operationalize formative evaluation in the instructional development process it became necessary to obtain author commitment to the principle by means of an intensive orientation and training program. It was difficult, however, to achieve a conceptual understanding of the entire process when dealing with a complex, multi-stage sophisticated model. Therefore, to begin the faculty-consultant dialogue and facilitate conceptualization of the process, a greatly simplified model was needed. After the process had been conceptualized, however, there was a need to explicate the various steps in the process so that procedures could be adapted to a specific case. At that time, a far more detailed model was required.

Since in the present study, in all three cases, both "mini" and "maxi" models were needed, it was concluded that several models of varying levels of detail should be available to carry out the process.

Heuristics

As a consequence of participating in the five field trials, the experimenter learned by successive discovery certain heuristics or rules of thumb, which may facilitate use of the model. Since these heuristics may be of value to those who might apply the MK II model or to other researchers in the field, they are presented at this time.

Heuristic 1: LISTEN TO THE STUDENTS; THEY ARE ONE OF THE BEST SOURCES OF INFORMATION FOR IDENTIFICATION AND REMEDIATION OF LEARNING PROBLEMS.

In three instances in which SLATEs were revised, the students provided strategic level solutions to major instructional problems. For SLATE A_1, students suggested a major reorganization and change of emphasis in order to clarify what was to be learned and to present it in what they perceived to be a more logical sequence. With reference to SLATE A_2, students suggested the deletion of a large amount of extraneous information which was hindering their learning of important content. They also proposed revision of related pre- and post-test items. In the case of SLATE B_1, students suggested teaching the concept of thousandths of an inch by developing the analogy that one penny is to $10.00 as 1/1000 is to one inch. This analogy provided an organizing structure to relate a number of disparate concepts. In short, the student groups provided unique and insightful solutions to their own learning problems--a skill which authors typically were unable to achieve because of their more sophisticated conceptualization of the subject matter.

Heuristic 2: INITIALLY TRY TO IDENTIFY AND REMEDIATE GROSS DISCREPANCIES IN THE PROTOTYPE.

The MK II model seems best suited for identification and remediation of the grossest, monumental types of discrepancies. MK II procedures

were primarily designed for a one-time effort since most authors or teachers did not have the time, desire, or resources for multiple iterative revisions. The "one shot" rationale led to the group debriefing structure which, in effect, generated a considerable amount of data in a very short time. The highly interactive and unstructured nature of the group produced information overload during the debriefings; so nuances of instructional problems were lost and only the major, or gross discrepancies were thoroughly conceptualized by the author. Nevertheless, remediation of these major problems made an enormous difference in the revised versions. In short, during the first prototype tryout with a group of nine to twelve students, the major learning problems emerged.

Heuristic 3: IF TECHNICAL ASSESSMENT AND REVISION HAVE NOT BEEN EFFECTIVELY CARRIED OUT, USE CAUTION IN PROCEEDING WITH STUDENT TRYOUTS.

The MK II procedures essentially involved a face-to-face feedback situation wherein a group of students evaluated and provided feedback directly to the author concerned. If the unit being evaluated was casually put together, unorganized, or technically poor, the student comments were so derogatory that the author became very embarrassed or defensive and the whole debriefing became an ego shattering disaster. This phenomenon seems most likely to happen to a novice SLATE designer who is precisely the one who can afford it least since he most needs feedback from students. The net effect may be that author becomes so humiliated that further development becomes difficult, if not impossible.

To avoid this unhappy contingency, it appeared that when employing MK II procedures, the prototype instructional units, like a prototype

aircraft, must be as carefully engineered and executed as humanly possible--preferably achieving some minimal level of sophistication prior to student tryouts. "Sophistication" in this sense means attention to technical details, organization, and continuity of the presentation. In short, do not tryout the prototype with a group of students until technical assessment is complete, and pedagogical, technical, or organizational details have been completely worked out.

Heuristic 4: FOREWARN AUTHORS THAT STUDENT CRITICISM CAN BE DEVASTATING, THEN USE MK II DEBRIEFING PROCEDURES ONLY WITH AUTHORS WHO HAVE A HIGH TOLERANCE FOR CRITICISM.

Students were blissfully unaware of the enormous effort required to develop a prototype SLATE. Consequently, when asked to criticize the product, they did so quite willingly if they perceived the author was genuinely interested and there would be no reprisals for telling it "the way it is." In providing their feedback, students were brutally frank, which meant the author had to listen while his product was critically dissected by a panel of judges. To maximize the interaction, it was necessary for the author to try and understand why the students encountered their problems rather than defend the unit. This was a difficult task unless the author made a conscious attempt to separate himself, as it were, from the fruits of his labor and accepted the criticism as it came. Authors who did not have a high tolerance for criticism tended to defend the unit rather than understand why the students had their problems; hence they were unable to remediate the difficulty.

Recommendations for Further Research

This study has raised a number of questions which are amenable to further research. These questions may be classified as: (1) improvements

or refinements to the model to make the tryout and revision process more effective and/or efficient; and (2) determining the generalizability of the model, e.g., whether the model in its present (or a different) configuration can be used for formative evaluation of other types of instructional systems. While these two types of questions interact, they will be treated separately for purposes of this discussion.

Research Leading to Refinements of the Model

In the context of formative evaluation of a specific self-instructional system, a number of confounding variables are operating which alter the nature of the feedback obtained regardless of the type of procedures specified in the model. In the area of group dynamics, for example, the group size, students' ability range, group objectives, and "ground rules" are specified in the model. Of these, changing the "ground rules" might be the most fruitful direction for further research. For example, one might look at alternate ways of structuring the group to meet the "needs" of the author desiring feedback. Clearly authors differ with respect to: (1) their ability to interact with small groups, (2) their ability to handle aversive feedback, and (3) how well the prototype was designed. All of these variables affect the nature of the debriefing feedback and raise such questions as: (1) should the author conduct the debriefing face-to-face, or should an author surrogate be used? (2) should authors be given desensitization and/or small group leadership training prior to tryout of their prototype instructional units? (3) how effective does a prototype have to be before the group debriefing procedures specified in the model are maximally useful?

With respect to the question of author in person versus author surrogate, it is possible that the surrogate would be a more objective data collector than the author himself. On the other hand, such an arrangement loses the impact of a face-to-face interaction between students and author. That is, authors may be able to disregard or misinterpret data collected by a surrogate easier than if they were personally confronted with the discrepancies.

Another series of questions which warrant research relate to the training of faculty to carry out formative evaluation independent of consultant help. What is the nature of a training program which enables faculty to revise their new instructional units prior to use with their classes? Is a tutorial training model necessary? Can the process be learned through simulation, or is first hand direct experience needed? How many training trials (revision of an instructional unit) are necessary for faculty of varying degrees of entering knowledge? How can aversive experiences which "turn off" faculty be avoided? How can the basic commitment of formative evaluation be obtained?

Related to the question of training, and in this experimenter's opinion a most fruitful area for research, would be to test this hypothesis: Training in and conduct of formative evaluation (e.g., feedback and revision) is an effective method of improving an author's basic design skills. Some evidence for improvement in design skills was observed in the present study, but the hypothesis warrants testing under more controlled conditions. It is possible that an effective way to teach design of instructional systems is simply teach formative evaluation and let each author evolve his own set of design heuristics based on student feedback.

Research Leading to Generalization of the Model

A much larger yet related domain of exploratory research are questions relating to generalizing the MK II procedures (or variations thereof) to instructional systems or components of increasing scope and complexity. The present study was restricted to a narrow part of the spectrum of possible types of instructional systems. Using the basic framework of the MK II model, exploratory research should be conducted to determine its generalizability to instructional system components such as lecture, laboratory, group discussion, or independent study. Beyond this, exploratory research is needed to determine how to operationalize the principle of formative evaluation within an instructional system such as a "course," a sequence of courses, a program, a curricula, a department, or a college. It is quite likely that the MK II procedures would have to be drastically modified to accommodate systems of varying scope and complexity.

A point of departure on the question of generalizability would be to focus the evaluation and revision efforts on the component which provides the majority of instruction to students (e.g., laboratory, lecture, SLATEs, discussion group, etc.). Although not the only component in the system, it is of central importance; i.e., if it fails, the system fails. Using this central component as the common experiential referent, one might utilize MK II procedures to obtain data not only on the central component but on the larger system as well.

The objective of a research program in generalizing the principle of formative evaluation would be to develop a tool kit of validated

alternative procedures which could then be incorporated into a training program for teaching design of instructional systems.

Assessment of the Model

In Chapter I, three criteria were stipulated for assessing the utility of the model: validity, feasibility, and effectiveness.

1. Validity.--Did the model do what it was supposed to do? In Chapter I it was stipulated that the model would be considered valid if (a) through its use the prototype lesson author was able to distinguish those sequences of instruction which were unsatisfactory, and (b) if the model predicted revision alternatives which remediated the unsatisfactory instructional sequences.

With respect to the criterion of validity, evidence in this study showed that the pattern of activities in the MK II model did facilitate identification of unsatisfactory instruction and allowed students (rather than the model itself) to predict strategies for remediating unsatisfactory instructional sequences. When authors implemented the strategies suggested by students, the data showed that student achievement and attitudes were significantly improved in two trials out of three. Therefore, it was determined that the MK II model was valid.

2. Feasibility.--The criterion stipulated in Chapter I was that the model would be considered feasible if fewer than twenty students were required for its use, and if faculty were willing and able to use it in the field situation.

With respect to the criterion of feasibility, evidence in this study showed that some faculty were willing and able to use MK II

procedures, but that for others the group confrontation was too threatening. For those willing to use the model, however, it appeared that they were able to do so after several training trials. Therefore, it was determined that the MK II model had conditional feasibility.

3. Effectiveness.--The criterion stipulated in Chapter I was that the model would be considered effective if comparative measures of student achievement and/or attitude between prototype and revised versions showed statistically significant differences in favor of the revised versions in 75% of the field experiments.

In the present study, prototype SLATEs were compared to revised versions in three different field experiments. Each experiment used four different measures. Thus, in the whole study, a total of twelve measures were used to assess differences between prototype and revised versions. Of those twelve measures, eight were statistically significant at the .01 level of confidence and one was statistically significant at the .05 level of confidence. Thus out of twelve measures, a total of nine were statistically significant; for an average of 75%. Since the MK II model did, in fact, lead to statistically significant differences in 75% of the measures used in the field experiments, the model was considered to have met the criterion of effectiveness.

Concluding Remarks

The MK II model provides an operational framework within which instructional development personnel can train or consult with faculty regarding formative evaluation of mediated self-instructional systems. Whether the model can be generalized to other types of instructional systems is a question yet to be answered. The general principles of the

model are as follows: (1) use a carefully developed prototype to provide a common instructional experience for a group of volunteer students of varying abilities; (2) collect data by means of learning and attitudinal measures after the common experience; (3) identify, discuss, and propose solutions to major problems by means of a group debriefing conducted by the author; (4) consult with "experts" on data interpretation; and (5) revise the instructional unit and recycle as necessary.

It is hoped that the principles developed in this study will prove useful to those in the field of instructional development and to other educators committed to improving our educational system.

BIBLIOGRAPHY

Alexander, L. T.; Kepner, C.; and Tregoe, B. "The Effectiveness of Knowledge of Results in a Military System Training Program." Journal of Applied Psychology, 46, 1962, 202-211.

Amidon, E., and Hunter, E. Improving Teaching. New York: Holt, Rinehart & Winston, 1966.

Anderson, Richard C. "The Comparative Field Experiment: An Illustration from High School Biology." Proceedings of the 1968 Invitational Conference on Testing Problems. Princeton: Educational Testing Service, 1969, 3-27.

Bany, M., and Johnson, L. Classroom Group Behavior. New York: Macmillan Co., 1964.

Barson, John. Instructional Systems Development: A Demonstration and Evaluation Project. Contract No. OE-5-16-025. East Lansing: Michigan State University, June, 1967.

Bradford, L.; Gibb, J.; and Benne K. T-Group Theory and Laboratory Method. New York: John Wiley & Sons, 1964.

Brethower, D. M., et al. Programed Learning: A Practicum. Ann Arbor: Ann Arbor Publishers, 1966.

Briggs, Leslie J. Handbook of Procedures for the Design of Instruction. Pittsburgh: American Institutes for Research, Monograph No. 4. 1970.

Briggs, Leslie J., et al. Instructional Media: A Procedure for the Design of Multi-Media Instruction, A Critical Review of Research, and Suggestions for Future Research. Pittsburgh: American Institutes for Research, Monograph No. 2, 1967.

Campbell, Donald T., and Stanley, Julian C. "Experimental and Quasi-Experimental Designs for Research on Teaching." Handbook of Research on Teaching. Edited by N. L. Gage. Chicago: Rand McNally & Co., 1963.

Chesler, M., and Fox, R. Role-Playing Methods in the Classroom. Chicago: Science Research Associates, 1966.

Davis, R. H. "SLATE Your Students for Structured Self-Tutoring." College and University Business. Vol. 44, No. 4, April, 1968, pp. 78-83.

Dick, Walter. "A Methodology for the Formative Evaluation of Instructional Materials." Journal of Educational Measurement, Vol. 5, 1968, 99-102.

Edwards, Allen L. Experimental Design in Psychological Research. New York: Rinehart & Company, Inc., 1950.

Flanagan, John C. "The Critical Incident Technique." Psychological Bulletin, LI, July, 1954, 327-358.

Fox, R., Luszki, M., and Schmuck, R. Diagnosing Classroom Learning Environments. Chicago: Science Research Associates, 1966.

Gilbert, T. F. "On the Relevance of Laboratory Investigation of Learning to Self-Instructional Programming." Teaching Machines and Programed Learning. Edited by A. A. Lumsdaine and R. L. Glaser. Washington, D.C.: NEA, 1960, 475-496.

Glaser, Robert. "Instructional Technology and the Measurement of Learning Outcomes: Some Questions." American Psychologist, Vol. 18, 1963, 519-521.

Glidewell, J. C.; Kantor, M.; Smith, L. M.; and Stringer, L. "Classroom Socialization and Social Structure." Review of Child Development Research. Edited by M. Hoffman and L. Hoffman. New York: Russell Sage Foundation, 1966, pp. 221-257.

Golembiewski, R. The Small Group. Chicago: University of Chicago Press, 1962.

Gropper, G. L.; Lumsdaine, A. A.; and Shipman, V. Improvement of Televised Instruction Based on Student Responses to Achievement Tests. Pittsburgh: American Institutes for Research, Report No. 7, Studies in Televised Instruction, 1961.

Hamreus, D. G. "Prototype Development." National Research Training Institute Manual. Monmouth: Teaching Research Division of the Oregon State System of Higher Education, 1969.

Hamreus, D. G. "The Systems Approach to Instructional Development." The Contribution of Behavioral Science to Instructional Technology. Monmouth: Teaching Research Division of Oregon State System of Higher Education, 1968.

Hare, A. P.; Borgatta, E. F.; and Bales, R. F. Small Groups: Studies in Social Interaction. New York: Alfred A. Knopf, 1955 and 1965.

Hare, A. P. Handbook of Small Group Research. New York: Free Press, a subsidiary of Crowell, Collier & Macmillan, 1962.

Henry, N. The Dynamics of Instructional Groups. Chicago: National Society for the Study of Education, 59th Yearbook, Part 2, 1960.

Horn, Robert E. Development Testing. Ann Arbor: Center for Programmed Learning for Business, 1966.

Kaplan, A. The Conduct of Inquiry: Methodology for Behavioral Science. San Francisco: Chandler, 1964.

Katz, D., and Kahn, R. The Social Psychology of Organizations. New York: John Wiley & Sons, 1966.

Kerlinger, F. N. Foundations of Behavioral Research. New York: Holt, Rinehart & Winston, 1964.

Likert, R. New Patterns of Management. New York: McGraw-Hill Book Co., 1961.

Lippitt, R.; Fox, R.; and Schmuck, R. Pupil-Teacher Adjustment and Mutual Adaptation in Creating Classroom Learning Environments. Final Report, U.S. Office of Education, Cooperative Research Project No. 1167, 1964.

Luft, J. Group Processes: An Introduction to Group Dynamics. Palo Alto, Calif.: National Press Books, 1963.

Mager, R. F. "On the Sequencing of Instructional Content." Psychology Reports, 9, 1961, 405-413.

Maier, Norman R. F. Problem Solving Discussion and Conferences: Leadership Methods and Skills. New York: McGraw-Hill Book Company, 1963.

McGrath, J., and Altman, I. Small Group Research. New York: Holt, Rinehart & Winston, 1966.

March, J., and Simon, H. Organizations. New York: John Wiley & Sons, 1958.

Markle, David G. The Development of the Bell System First Aid and Personal Safety Course. Palo Alto: American Institutes for Research, 1967.

Markle, Susan M. "Empirical Testing of Programs" (Chapt. V). Programmed Instruction. Sixty-sixth Yearbook of the National Society for the Study of Education, Part II. Edited by Phil C. Lange. Chicago: University of Chicago Press, 1967.

Olmstead, M. The Small Group. New York: Random House, 1959.

Paulson, Casper F. "Evaluation of Instructional Systems." National Research Training Manual. Edited by Jack Crawford. Monmouth: Teaching Research, Division of the Oregon State System of Higher Education, 1969.

Pipe, Peter. Practical Programming. New York: Holt, Rinehart & Winston, 1966.

Popham, W. James. "Curriculum Materials." Review of Educational Research, XXXIX, June, 1969, 319-338.

Raven, B. Bibliography of Publications Relating to the Small Group. Technical Report, No. 1. Office of Naval Research, 1959.

Robeck, M. A Study of the Revision Process in Programed Instruction. M.A. Thesis, UCLA, 1965.

Rothkopf, E. Z. "Learning from Written Materials: An Exploration of the Control of Inspection Behavior by Test-Like Events." American Educational Research Journal, Vol. 3, 1966, pp. 241-249.

Rothkopf, E. Z. "Two Scientific Approaches to the Management of Instruction." Learning Research and School Subjects. Edited by R. M. Gagne and W. J. Gephart. Itasca, Illinois: Peacock Publishers, 1968, pp. 107-149.

Schalock, H. D. "Measurement." National Research Training Manual. 2nd ed. Edited by Jack Crawford. Monmouth, Oregon: Teaching Research Division of the Oregon State System of Higher Education, 1969.

Schein, E., and Bennis, W. Personal and Organizational Change Through Group Methods. New York: John Wiley & Sons, 1965.

Schmuck, R.; Chesler, M.; and Lippitt, R. Problem Solving to Improve Classroom Learning. Chicago: Science Research Associates, 1966.

Schmuck, Richard A., and Schmuck, Patricia A. Group Processes in the Classroom. Dubuque: William C. Brown Co., 1971.

Schutz, Richard E. "Experimentation Relating to Formative Evaluation." Research and Development Strategies in Theory Refinement and Educational Improvement. Madison: University of Wisconsin Press, Center for Cognitive Learning, 1967.

Scott, R. O., and Yelon, S. L. "The Student as a Co-Author--The First Step in Formative Evaluation." Educational Technology, Vol. 9, No. 10, 76-78.

Scriven, Michael. "The Methodology of Evaluation." Perspectives of Curriculum Evaluation. Edited by R. W. Tyler, R. M. Gagne, and M. Scriven. AERA Monograph Series on Curriculum Evaluation, Vol. 1. Chicago: Rand McNally, 1967.

Shepherd, C. Small Groups. San Francisco: Chandler Publishing Co., 1964.

Short, J. G., et al. "Strategies of Training Development." Report No. AIR-E97-2/68. Palo Alto: American Institutes for Research, 1968.

Silberman, H., and Coulson, J. Use of Exploratory Research and Individual Tutoring Techniques for the Development of Programming Methods and Theory. Santa Monica: System Development Corp., June, 1965.

Silvern, Leonard C. "LOGOS: A System Language for Flowchart Modeling." Educational Technology, IX, June, 1969, 18-23.

Skinner, B. F. "The Science of Learning and the Art of Teaching." Current Trends in Psychology and the Behavioral Sciences. Pittsburgh: University of Pittsburgh Press, 1954.

Skinner, B. F. "Teaching Machines." Science, Vol. 128, October 24, 1958, 969-977.

Smith, Robert G. The Design of Instructional Systems. Alexandria: The George Washington University Human Resources Research Office, 1966.

Taber, Julian I.; Glaser, Robert; Schaefer, Halmuth N. Learning and Programmed Instruction. Reading: Addison-Wesley, 1965.

Vandermeer, A. W. An Investigation of the Improvement of Educational Filmstrips and a Derivation of Principles Relating to the Effectiveness of these Media. University Park, Penn.: Pennsylvania State University Press, 1964, 154.

Vandermeer, A. W., et al. An Investigation of the Improvement of Educational Motion Pictures and a Derivation of Principles Relating to the Effectiveness of these Media. NDEA Report No. VIIA-225. University Park, Penn.: Pennsylvania State University. ERIC #ED 003 536. Washington, D.C.: USOE, April, 1965, 95.

Wiener, Norbert. The Human Use of Human Beings. New York: Doubleday and Company, 1954.

APPENDIX A

SLATE AUTHOR

INTERVIEW QUESTIONNAIRE

SLATE AUTHOR
INTERVIEW QUESTIONNAIRE

DATE _____

NAME _____ AFFILIATION _____

EXPERIENCE IN PI DESIGN _____

1.0 Describe the subject matter and target population and instructional system in which your materials are being used.

2.0 Did you revise the materials after a "first draft" or prototype had been produced?

3.0 If so, what was your general strategy for the revision aspect of program development? Did you have a predetermined strategy, or did the revisions just evolve randomly?

3.1 If you used a formal plan, what was it? Had you used it before?

4.0 From whom did you obtain feedback--individual students, groups, "experts?"

4.1 What was your selection criteria? What were the critical attributes of the people you selected for tryout? Previous skills, subject matter competence, availability, intelligence, volunteers, etc.

5.0 What kinds of feedback did you try to get?

5.1 Achievement data

5.2 Attitudinal data

5.3 Process data--what were the hangups or problems during the tryout?

5.4 Background data--assessment of entry skills, prerequisites, demographic data?

6.0 How did you gather the various kinds of feedback data? What techniques and procedures were used?

6.1 Achievement data--objective tests--how validated, reliability, etc.?

6.2 Attitudinal data--rating scale, verbal interview, etc.?

6.3 Background data?

6.4 Process data--an observer watched them or what?

6.4.1 If using an observer, what help, if any, was given students? What general procedures were used? How was the observer selected? Trained?

7.0 When did you collect the data, early or late in the development? Before and/or after first draft (AV and print) was complete?

8.0 Did your data help you identify the nature of student problems? How would you classify the type of problems you found?

8.1 Administrative/technical

8.2 Communication variable

8.3 Learning variable

Analyze for gap, mastery and irrelevancy.

9.0 How did you score, display and analyze data--use item analysis, matrix, consultants for data interpretation?

10.0 How did you determine if a revision was really necessary? How much of what type of data was the "critical mass" which said--revision is mandatory?

11.0 Did you weigh revising the objectives, the evaluation instrument, or center only on revising the print and AV materials?

12.0 How did you determine what would be the design of the revised materials? Did you have some model, ad hoc, expert advice, or your own experience?

13.0 How many revisions did you make? Did the techniques of data collection, analysis and design change for each revision or did you use pretty much the same procedures throughout?

14.0 What do you believe are the critical steps in the tryout and revision process; e.g., select the "right" students, watch them closely, or use large groups?

15.0 How much time and effort did you spend on revision? Percent of original prototype development time? (guess)

16.0 Was it worth it? How did you decide how much better or worse are revised materials than the original?

17.0 If you had to do it all over, how would you do it differently? Would you incorporate student feedback earlier in the process? Later? Different technique?

APPENDIX B

EXPLANATION OF STEPS

IN THE

MK II "MAXI" MODEL

EXPLANATION OF STEPS IN

THE MK II "MAXI" MODEL

Description of the MK II "Maxi" Model

Each of the five functions in the "mini" version are made up of a number of constituent tasks. Each task, in turn, has component subtasks, which orchestrate together to perform a given function. The functions coordinate to perform the overall task of formative evaluation. Instructional Development consultants must know each constituent task and the specific techniques to be applied. Therefore, the detailed MK II "maxi" model was developed including subtasks down to the fourth level of detail. The following explanation relates to each step in the "maxi" model as shown in Figure 8 in Chapter III.

Box 0.0 Enter With Prototype

To begin formative evaluation, one must have a fully developed prototype. "Prototype" connotes that all of the instructional materials have been completed without having obtained feedback from technical experts, or students from the target population. "Prototype" also denotes that evaluation instruments have been designed based on specific learning objectives and that all instructional components are ready for student use. The prototype is not a rough draft, quickly put together. It is a finished product, as capable of performing its instructional function as the designer's capabilities will permit.

Box 1.1 Prepare for Consultant Tryouts

Assuming a prototype is ready to be evaluated, the first set of

activities relate to logistical preparation prior to obtaining feedback on technical problems.

Box 1.1.1 Determine data format.--The author must determine whether he wants feedback on technical problems in the form of a verbal debriefing, a written report, a rating scale, questionnaire, or other device. It is strongly suggested that a standard format be adopted such as a "storyboard" upon which discrepancies can be noted as they occur. A storyboard is essentially a verbatim script together with simple drawings of all visuals. The storyboard, together with SLATE workbook, and evaluation instruments provide a standard format for the consultant to make notes.

Notes should be augmented by a face-to-face debriefing between consultant and author. Such an interaction serves to facilitate author understanding of the discrepancies and expedites arriving at a solution; e.g., the consultant can provide input on how to remediate discrepancies.

The suggested data format, therefore, includes both a written narrative such as marginal notes on the storyboard, workbook, and/or evaluation instruments, as well as verbal debriefing between the SLATE author and consultant.

Box 1.1.2 Select and brief consultants.--Three types of consultants are required; subject matter expert, media specialist, and evaluation specialist. The most important characteristic of these consultants is that they are perceived by the SLATE author as highly credible sources of information. Since perceived credibility is a subjective judgment, little guidance can be provided relative to consultant selection.

Assuming that credible experts are available, the SLATE author should brief them on: (1) the type of information required (Box 2.0); (2) format to use--including explanation of the use of the storyboard for notetaking and need for a verbal debriefing/problem solving session; (3) when the feedback is needed--the time frame within which the consultant feedback must be obtained.

Box 1.1.3 Reproduce and distribute materials.--The final step in preparation for consultant tryouts is to reproduce the prototype materials, storyboard, and instruments for distribution to the selected consultants.

Step 2.0 Collect Technical Review Data

At this time, each of the consultants interact with the prototype materials and identifies discrepancies in their respective interest area.

Box 2.1 Subject Matter

Since naive students can learn inaccurate and out-of-date information as readily as more desirable content, the subject matter expert should assess all SLATE materials (slides, tapes, films, models, workbooks, problems, specimens, experiments, etc.)for TECHNICAL ACCURACY (Box 2.1.1) and UP-TO-DATENESS (Box 2.1.2). TECHNICAL ACCURACY, for example, refers to an assessment of the propositions, assertions, inferences, assumptions, evidence, and judgments in the SLATE presentation, as compared to other statements made previously by others knowledgeable in the discipline. Such mundane things as spelling and correct use of technical terms, as well as a more professional assessment of the overall accuracy of the presentation should be accomplished. UP-TO-DATENESS (Box 2.1.2), would reflect an assessment as to whether the propositions, inferences, evidence, etc., reflect the latest developments in the discipline.

The two other factors which should be considered by the subject matter expert are both subjective in nature and do not necessarily mandate a revision. One such factor is the overall TREATMENT of the subject matter (Box 2.1.3). The term TREATMENT refers to a diffuse number of stylistic variables, such as organization (inductive or deductive), use of humor, satire, redundancy, novelty, use of advanced organizers and summaries. The assessment of TREATMENT essentially asks the question: Did I enjoy it--would students enjoy it? Is it too didactic, conversational, flippant, or disorganized?

It should be reiterated that unfavorable assessment of TREATMENT and/or OBJECTIVES does not constitute a mandate to revise, until corroborated by students. However, discrepancies in TECHNICAL ACCURACY and UP-TO-DATENESS should be remediated before tryouts with students.

Box 2.2 Instructional Media Quality

The media specialist is, therefore, asked to assess the SLATE instructional stimuli and carrel environment, not to produce a technical masterpiece, but rather to identify and eliminate technical problems which may affect student learning. In this context, the media specialist is asked to assess five factors:

1. Audio quality of tape recorded information (Box 2.2.1) to identify gross discrepancies in audibility or excessive background noise.

2. Video quality of film, slides, photographic prints or other visual stimuli, to identify gross discrepancies in focus, exposure, legibility or printed data in the visuals.

3. An assessment of the legibility and format convenience of all print materials, such as workbook, handouts, directions, etc.

4. Technical quality of any other instructional media used, such as models, specimens.

5. Assess the carrel and its component equipment for reliability and for sufficient screen illumination.

Once identified by a competent media specialist, these potential impediments to learning can often be remediated without duplicating the original production effort. Furthermore, if such deficiencies are not removed prior to student tryouts, feedback from students may focus on these relatively obvious technical problems, attentuating feedback on critical learning variables, such as language, rate, sequence, or practice.

Box 2.3 Evaluation Instrument Quality

The prototype SLATE, like a prototype aircraft must be highly instrumented during its initial "test flights" to provide a maximum amount of relevant data back to the designer. After sufficient data have been collected and modifications made so the SLATE (or aircraft) performs more closely to the design specifications, then much of the instrumentation can be removed on the "production" models.

While the concept of "overinstrumentation" of the prototype has validity in the aircraft industry, there is a great deal of reluctance to apply this concept to the instrumentation of prototype SLATEs. The reason is simple. The SLATE author, unlike an aircraft company, does not have a huge engineering staff to develop instruments for his new vehicle. Instead, virtually every instrument, whether it be to assess learning, attitudes, or prerequisite skills, must be designed and developed by the author. This resultant increase in workload, coupled with increases already assumed as a result of planning and production of the SLATE, is simply too much in many cases. The result is that the critical instrumentation/evaluation function is often neglected.

While most ISD process models suggest that evaluation instruments be developed early in the process, e.g., simultaneous with formulating lesson objectives--personal experience has shown that this rarely happens. Most often, evaluation instruments are developed after the "content" has been identified and may or may not reflect the lesson objectives. Moreover, many SLATEs may not be designed with any evaluation instruments at all because the SLATE is embedded into a larger system (a course) and is not evaluated as an individual entity, but as part of the larger system. Simply stated, unless a number of evaluation instruments are embedded into the prototype SLATE, formative evaluation becomes unsystematic, unscientific, and practically impossible.

For these reasons, a check must be made by a competent evaluation specialist or instructional technologist for blatant deficiencies in the prototype SLATE evaluation instruments. Such deficiencies include:

1. Omission of critical instruments.
2. Omission of a student performance criterion.
3. Non-correspondence between evaluation instruments and SLATE objectives.

For purposes of instrumentation of a prototype SLATE, four types of instruments are considered critical. These four are: pre-test, post-test, en route tests, and attitudinal measures.

Box 2.3.1 Pre-tests.--Pre-tests may be of two types: either to assess prerequisite competencies or to establish a benchmark against which later learning may be assessed. In the case of the former, the test is essentially a screening device in that some minimal score is required before access to the SLATE is permitted. On the other hand, the "benchmark" type of pre-test is one in which the students' entering

subject matter competencies are assessed for later comparison against performance on an equivalent form post-test.

Box 2.3.2 Post-test.--Essentially an equivalent form of the "benchmark" pre-test. The pre- and post-tests should both reflect evaluation of competencies stated in the SLATE objectives and measure student performance against an absolute standard instead of comparing students one against the other.

Box 2.3.3 En route tests.--En route tests are simply techniques for systematizing responses and feedback to the students so they can assess their own progress and for structuring the learning environment so that students do not progress to high level competencies until pre-requisites are "mastered." While all subject matter does not necessarily lend itself to a hierarchically organized presentation, all learners do require feedback to assess their progress. Furthermore, en route tests will help pinpoint the exact place in a SLATE which is causing a learning problem much like frame analysis in programmed instruction.

En route tests may use a variety of formats such as "frames" of constructed responses, problems, or open-ended questions. The point is, their major functions are: (1) to evoke responses and provide feedback to students and (2) through analysis of student errors on pre-, post-, and en route tests, help pinpoint learning problems in the prototype.

Box 2.3.4 Attitudinal measures.--Student attitudes towards the subject matter and instructional techniques have gained increasing recognition as important outcomes of instruction. Seldom, however, have affective outcomes been assessed as an integral part of formative evaluation. Instead, "end-of-term" summative evaluation surveys are used which

are so general that specific problems are difficult to identify. If a prototype SLATE "turns off" students, the author should become aware of this problem and take remedial action before it is used with large numbers of students.

For these reasons, the MK II model stipulates the use of an attitude survey, or "reactionnaire" as an integral component of prototype instrumentation. After several iterations of attitudinal data have been gathered and appropriate modifications made in the prototype, the frequency of use of the attitudinal instrument may be diminished to reduce the logistics of operating the "production" SLATE. It is important, however, that the evaluation specialist include an assessment of this instrument during the technical review (Box 2.3.4).

It is to be expected that few, if any, SLATE authors will have developed an adequate attitude instrument. Therefore, the experimenter developed a twenty-seven item attitude survey instrument (Appendix G) which may be used in cases where no other instrument exists or where the existing instrument is judged to be inadequate.

In summary, Step 2.0, COLLECT TECHNICAL REVIEW DATA, includes assessment of the prototype SLATE instructional stimuli from the standpoint of technical accuracy of content, technical quality of instructional media, and inclusion of the four types of evaluation instruments.

Step 3.0 Collect Student Tryout Data

This step is performed chronologically after one complete cycle in which revisions were developed based on the TECHNICAL REVIEW data. Note that the RECYCLE (Step 6.0) sends the process back through a second iteration in which data is collected, analyzed, and revisions developed

based on STUDENT TRYOUT DATA (Step 3.0). Therefore, in this appendix, both Box 1.3 (PREPARE FOR STUDENT TRYOUTS) and Step 3.0 (COLLECT STUDENT TRYOUT DATA) are discussed after Step 6.0 (RECYCLE).

Step 4.0 Analyze Data

After feedback data has been collected, the next task is to analyze the discrepancies.

First, the author lists the deficiencies identified in order of priority; then conducts a tradeoff analysis to determine which deficiencies warrant revision in light of the available resources and seriousness of the deficiencies. Basically, a go-no-go decision must be made for each discrepancy noted in the priority listing. That is, for each discrepancy identified a decision must be made to revise or not to revise.

If the decision is made to revise; then the process moves on to the DEVELOP REVISIONS step in Box 5.0. However, if resources, time, or other constraints preclude remediating a given discrepancy, the process moves instead to the RECYCLE step in Box 6.0 which asks the question: Is additional feedback and revision warranted before stipulating the SLATE is ready for operational use with the target population?

Box 4.1 List Deficiencies in Rank Order

In order to make the go-no-go decision, the first activity is to list all discrepancies revealed by previous data collection in order of their seriousness. During cycle one, any problems identified by the technical experts must be synthesized into a master priority listing. In cycle two the deficiencies listed will be those uncovered during student tryouts and debriefings.

For example, a typical discrepancy might be that 80% of the students missed a certain item on the post-test, while another discrepancy might be that 30% of the students indicated on a questionnaire that they were "not aware of the SLATE objectives." These two discrepancies, along with all the others identified by the student tryout group should be placed in hierarchical order, with the most serious discrepancy given top priority.

Box 4.2 Tradeoff Analysis

After the problems are listed in order of their seriousness, a TRADEOFF ANALYSIS is conducted (Step 4.2). The factors or criteria to be considered in the TRADEOFF ANALYSIS are stipulated in boxes 4.2.1 through 4.2.5. Each of these factors represents a heuristic consideration which will affect the go-no-go decision.

For example, if the most serious discrepancy in a prototype SLATE is that 80% of the students missed a certain item on the post-test; then assessment of CAUSAL FACTORS (Box 4.2.1) would begin the TRADEOFF ANALYSIS. Determination of CAUSAL FACTORS is prerequisite to the development of a FEASIBLE REVISION HYPOTHESIS (Box 4.2.4).

Analysis of CAUSAL FACTORS is both extremely critical and difficult. There are no algorithmic "trouble shooting" flowcharts which state precise cause and effect. It is extremely important, therefore, that during the student or consultant debriefing the SLATE author ascertain the underlying WHY behind each problem; e.g., ascertain the causal factors as much as possible from the best source of information possible--the students or consultants concerned.

If CAUSAL FACTORS were not pinpointed during the face-to-face feedback with students (or experts), then post hoc analyses must be

employed as a last resort. Such post hoc analyses involve the development of a student-test item matrix (Appendix E) with columns representing test items and rows representing students. An X indicates a correct response.

The first question to be answered during post hoc analysis of CAUSAL FACTORS is how "good" were the test items. It is possible that a missed post-test item indicates a lack of student capability due to lack of practice, inattention, or poor message design. It is also possible that the student had achieved the capability but the particular test item was ambiguous or the score key wrong.

A conventional item analysis can contribute to decisions regarding quality of test items. For example, an item would be "poor" (and should be revised) if a large proportion of the lower ability students answered correctly while a large proportion of high ability students answered wrong.

On the other hand, if a test item is determined "good," although it was still missed by 30% or more of the tryout group, then the causal factors would presumably be related to either: (1) message design, (2) insufficient practice, (3) inattention, or (4) some other unknown factor.

If an en route response corresponding to the missed item was correct, then it may be presumed that lack of practice may be the relevant variable (Baker, 1970), and revisions could be designed accordingly. If the en route response and the corresponding post-test item were both wrong, then it may be assumed that inattention, poor message design, or some other unknown factor is responsible. In these latter cases, the development of revisions in the absence of direct comfirmatory feedback

from students becomes sheer specualtion on the part of the designer and probably a waste of time.

In short, assessment of CAUSAL FACTORS must be done largely at the time the problems were identified, when the author is talking face-to-face with the students. After this, identification of CAUSAL FACTORS other than "poor" test items becomes so complex that revisions become sheer speculation. Under these circumstances, it is advisable to make a "no-go" decision and RECYCLE, e.g., begin again to collect more data.

Box 4.2.2 Number of components affected.--Assuming the causal factors for the top priority problems can be reasonably inferred, the next step in TRADEOFF ANALYSIS is to assess the NUMBER OF COMPONENTS AFFECTED. This factor refers to the degree of interrelatedness of the instructional stimuli which are the presumed cause of a discrepancy, e.g., the slides and tape, the tape only, workbook only, etc. Obviously, the greater the number of components involved, the greater the cost and effort to remediate the problem. For example, an overly technical or ambiguous graph contained in the workbook may be remediated by redesign and retyping on mimeo stencil. On the other hand, if this same graph is shown on a slide and is discussed on the audio tape, considerably more revision effort is required.

Box 4.2.3 Time and resources available.--A third factor in the tradeoff analysis is TIME AND RESOURCES AVAILABLE. If there are zero time or resources available, a "no-go" decision is mandated. On the other hand, if limited time and resources are available, each "go" decision must maximize payoff. For example, an evaluator may have to determine whether it is better to remediate a dozen small discrepancies or one major one.

Box 4.2.4 Feasible revision hypotheses.--The FEASIBLE REVISION HYPOTHESES step asks the formative evaluator to consider the range of variables stipulated in 5.0 and develop a preliminary hypothesis as to what appears most feasible to solve a given discrepancy in light of the causal factors and operating constraints. This asks: given the nature of the discrepancies involved (e.g., causal factors) what will fix them? what are feasible alternative solutions to a specific problem?

To facilitate awareness of a number of alternatives, several variables which may be manipulated to solve specific problems are listed in Step 5.0, DEVELOP REVISIONS. The hypothesis actually selected will depend on a number of interactive factors such as: (1) causal factors, (2) priority of the discrepancy, (3) theoretic power of the alternative to solve the problem, (4) cost of the alternative, and (5) resources available.

Having made a preliminary decision as to HOW a discrepancy may be remediated, the evaluator must now determine whether it is worth remediating.

Box 4.2.5 Estimated cost of revision.--For the revision hypothesis selected, the cost must be calculated and weighted against the RESOURCES AVAILABLE so a final go-no-go decision can be made.

In sum, the TRADEOFF ANALYSIS asks the formative evaluator to rank order the problems, assess the probable causes, and select a feasible solution within the constraints of his resources and ability. For those problems and solutions thus selected, a decision is made to "go," to commit additional resources to remediation, and the process enters the DEVELOP REVISION stage. For those problems which did not warrant revistion, e.g., a "no-go" decision, the process enters the RECYCLE step

(Box 6.0) which asks the question: Is the SLATE operationally ready or is additional feedback warranted?

Step 5.0 Develop Revisions

At this point in the process, the formative evaluator must DEVELOP REVISIONS for those discrepancies which warranted a "go" decision. Step 5.0 contains a number of revision variables which function as a "cafeteria" for development of revision hypotheses. Due to time and space limitations, it is not possible in the context of this study to develop a set of if-then heuristic statements relative to design of revisions. Instead, the variables considered by this author to be critical in instructional design have simply been described so that users of the model will recognize the range of alternatives available. It is suggested that at this stage of the process the author of the prototype obtain consultant help in design of revisions.

Box 5.1 Message Design

It is likely that the most frequently selected revision hypothesis will involve some form of message revision. The term "message" is used here in a broad sense to include any auditory or visual stimulation eminating from the presentation media, exclusive of the evaluation instruments. In other words, the SLATE is regarded as an ongoing "message" except during administration of pre-, post-, en route, or attitudinal instruments. Conceived in this way, the "message" has several classes of variables which can be manipulated to facilitate communication.

Box 5.1.1 Content.--Revision of CONTENT relates to any changes in facts, propositions, assertions, evidence, inferences, or examples stated in the audio or visual instructional stimuli. Major additions

or deletions of content must be reflected in evaluation instruments and lesson objectives; whereas, additional examples or clarification of existing propositions may be accomplished independent of exams and objectives.

Box 5.1.2 Treatment.--Treatment relates to the overall style of presentation, e.g., inductive, deductive, humorous, satiric, or expository. One may have found that students were bored during the presentation. One alternative would be to change the treatment from deductive/expository to inductive/humorous.

Box 5.1.3. Organization/sequence.--Research by Gagne (1965) and Briggs (1968) has shown that sequencing information in a hierarchical order can facilitate acquisition and retention. However, research by Mager (1962) and Fry (1970) indicated that organization and sequence interact with learning styles or other individual differences so that an important source of redesign information would be the students at the debriefing.

Box 5.1.4 Message complexity.--Human learners, like computers, have a limited information processing capacity. The audio and visual stimuli used to present information can easily overload the learner's processing ability, resulting in a decrement in learning. The variable "message complexity" relates to the amount of information presented to the learner/per unit of time. In general, the more complex the message, the greater will be the strain on the students' processing mechanisms. There is some evidence that for cognitive learning objectives, messages of relatively low complexity are appropriate (Travers, 1964). However for objectives in the affective domain highly complex messages may be

most successful (Perrin, 1970). Some dimensions of MESSAGE COMPLEXITY are as follows.

Box 5.1.4.1 Sense modality. Refers to whether the audio and visual sense modalities are used simultaneously or sequentially. Complexity, of course, is increased when using two modalities simultaneously.

Box 5.1.4.2 Redundancy. Refers to repetition of an idea within a sense modality, or across modalities. For example, English language has been shown to be approximately 50% redundant. If these same ideas are expressed visually as well as orally, the redundancy is further increased.

Box 5.1.4.3 Word/picture relationships. Refers to the "relatedness" and "dominance" of audio and visual stimuli which are juxtaposed, e.g., occuring at the same time. There is some evidence that either words or pictures may dominate or provide the majority of information to a given learner on a given task (May, 1965). However, at present there is no way of predicting what mode (words or pictures) will or ought to dominate. "Relatedness" appears to be a subjective judgment regarding the idealogical similarity between words and pictures. Some evidence indicates that to optimize cognitive learning tasks, words and pictures should quality or explain each other (May, 1965).

Box 5.1.4.4 Rate of presentation. Rate of presentation may be qualified as words per minute or visuals per minute, but this does not reflect the language difficulty level, the visual complexity, or the idealogical content of the message. Rate, like other variables, is not absolute but relative to a given set of learners on a given task. Hence the best source of information with regard to rate are the students at the tryout sessions.

Box 5.2 Revise Student Response and/or Feedback

Research in programmed instruction has long since demonstrated the necessity for student responses, either overt or covert. Responses alone, however, are not sufficient unless accompanied by "feedback" or knowledge of results. Research using non-programmed information has also corroborated the desirability of incorporating student response and knowledge of results (Rothkopf, 1965). Consequently, one of the major alternatives in development of revision hypotheses for SLATEs is the domain of student response and feedback. Both response and feedback have three dimensions: frequency, format, and type.

Box 5.2.1 Frequency. Frequency refers to how often a response is evoked and feedback provided. Research on this point is equivocal, since frequency seems to be a function of the task, learner individual differences, and the theoretic biases of the researcher. In general, however, responses and feedback should be frequent enough so that the learner is aware of his progress and deficiencies. Again, the student debriefing is the ideal source of information to determine optimal response/feedback frequency.

Box 5.2.2 Format. A response and feedback can be accomplished in a number of ways: erasing answer sheets, write-in, multiple choice questions, or a motor performance. If students are having difficulty with a particular response, it may be an unfamiliar format rather than lack of capability which is causing the problem. If such is the case, the form of the response may be revised.

Box 5.2.3 Type. Responses may be classified as "enabling" or "criterion." The former, of course, is designed to allow student practice

of component learning tasks. The latter is to assess (and provide feedback) on the achievement of specified learning tasks. Success on "enabling" responses followed by failure on "criterion" responses indicates insufficient practice.

Step 6.0 Recycle

Assuming that some discrepancies have been revised and others not revised, there is a need to first INTEGRATE REVISED AND UNREVISED COMPONENTS INTO A NEW SLATE (Box 6.1). That is, the old and new must be organized to form a new, unified SLATE. Following this, the formative evaluator must DETERMINE IF ADDITIONAL FEEDBACK IS WARRANTED. This decision is predicated on how closely the SLATE, as revised, is expected or hypothesized to achieve stated objectives. If, for example, the desired operationally ready criteria is the famous 90/90 criteria where 90% of the students will achieve 90% of the objectives and if the SLATE prior to revision was operating at a 50/50 level, it is highly doubtful whether even major revisions will achieve the desired improvement on the first revision. Logically there is likely to be a need for further feedback and the decision would be made to recycle through the data collection and revision steps again. On the other hand, if the SLATE performed fairly close to criterion (e.g., 70/70), then it is reasonable to assume that the revisions which were accomplished may achieve the 90/90 level, and no further feedback is absolutely required. If possible, it is desirable to continue to RECYCLE until the resources have been expended or until the SLATE achieves the desired criterion, whichever comes first.

Assuming the technical review has been completed and revisions developed accordingly, the next step in the process is to RECYCLE back to Step 1.3 and begin to PREPARE FOR STUDENT TRYOUTS. Following completion of the preparation steps, the process moves on to Step 3.0, COLLECT TRYOUT DATA. After data from student tryouts are collected, the process follows the same sequence of steps presented earlier: Step 4.0 ANALYZE DATA, Step 5.0 DEVELOP REVISIONS, and Step 6.0 RECYCLE.

Since a detailed breakdown of the operations and rationale for boxes 1.3 and 3.0 are provided in Chapter II and Chapter V, they will not be reiterated in this appendix.

APPENDIX C

"AGENDA" FOR

MK II TRYOUT/DEBRIEFING

"AGENDA" FOR
MK II TRYOUT/DEBRIEFING

Instructional Development Tryout Session

I. Preflight Facility:

Check software installation and operation in each carrel. Check for required number of workbooks, pre- and post-tests, answer sheets, keys, data matrices, reactionnaires, audio recording equipment and problem posting flip chart, and refreshments.

II. Student Arrival:

1. Pass out name tags

2. Create atmosphere of informality and low threat

Students have volunteered for this session and are unsure as to whether this will adversely affect their grade in the course, future employment, or other more horrible reprisals. They must be put at ease or very little constructive criticism will be forthcoming. Therefore, wear informal clothes (the student will) and make small talk as students arrive.

III. Introductory Remarks:

1. Welcome:

Thank students for their willingness to help you revise your "first draft" materials. Assure them that their frank and honest opinions are of crucial importance and that nothing they say will in any way affect their grade, job, or pose other threats. It is the author and the program which is under the gun--not them.

2. Role of Students:

To help you identify weaknesses in the materials, procedures, or exams, and to make comments and/or suggestions for improvement. You are looking for comments pro and con on "relevance," "redundancy," "boredom," "obscurity," "clarity of visuals," "needless make work," poor exam questions, etc.

3. Role of Author:

Your role is to gather data and suggestions for revising the materials and to provide tutorial assistance to the students on any aspect of the lesson.

4. Overview of the Procedure:

The tryout will begin with a pre-test (to assess how much they know to start with); then use the lesson materials; then a post-test (to determine how much they have learned from the materials); followed by an opinionnaire and then a break, with refreshments. After the break will be a group debriefing.

IV. General Instructions:

1. Test Scoring: Both pre-test and post-tests are self-scoring; students score their own. Please mark incorrect answers on the answer key--not in the test booklet.

Scores do not count towards a grade; they are for your information and to show us weaknesses in the lesson.

2. Be Honest: Don't look at the answer key before or during the exams. If you artificially inflate your score, we don't really know how good (or bad) the lesson is.

3. Guessing: Guess at the answers you don't know, and place a question mark after your answer on the test booklet. If you don't understand the question, place a question mark in front of the question in the test booklet and the answer key.

4. Ask for Help: If you have problems during the lesson, raise your hand and I will come over. Do not talk to your neighbor.

5. Write Down Your Problems: When you have a problem, write it down in the workbook.

6. Reactionnaire: We need your opinion on several critical aspects of the lesson design. Be frank and honest as you fill this out.

7. Break: Have a coke and donut and don't go away. We need you for the debriefing.

8. Debriefing: We will reconvene to discuss the lesson, using exam scores, reactionnaire data, and your notes and comments to organize the discussion. Remember, any comments you make will be useful.

APPENDIX D

"CHECKLIST" FOR MK II

TRYOUT AND DEBRIEFING

"CHECKLIST" FOR MK II

TRYOUT AND DEBRIEFING

AH 111 Instructional Development SLATE Tryout Procedure 21 October 1970

AGENDA

I. INTRODUCTORY REMARKS

1. Welcome: thank students for their willingness to participate in the tryout.

2. Introduction: doctoral research experimenter and AH grad assistants.

3. Name Tags: pass out name tags and explain they will help identification throughout the session.

4. Role of Student: to help designer identify weaknesses in the set of new materials. Comments and suggestions WILL be utilized for revisions.

5. Overview of Procedure:

   a. Pre-test: We must find how much you already know about the subject matter to determine how much you have learned tonight and see how good or bad the materials are--hence the pre-test.

   b. Sure or Unsure Measure: we need to know if you "really know" something or if you were a good guesser. Circle S or U on tests.

   c. Take the Program: again reiterate it is the materials not the students being evaluated.

   d. During the Program: designer will circulate to answer questions.

      1. Do not talk to each other--ask the designer.

      2. Write your comments/questions in the margin of the workbook "not clear," "too fast," "irrelevant," "busywork," etc.

      3. Raise your hand and designer will come to you.

      THESE COMMENTS AND QUESTIONS ARE CRITICAL--SO DON'T BE SHY

      4. You may smoke, or take a short (1-2 min.) break when you want to.

e. Post-test: same as the pre-test, and will give us a measure of the teaching effectiveness of the materials--weaknesses.

f. Reactionnaire: immediately after post-test, while your memory is fresh, answer several questions about how you felt about important design aspects of the materials.

g. Break: 15 minute, coffee and coke, donuts supplied by the house.

h. Debriefing: very critical discussion following the break to explore your questions and comments, and obtain your recommendations on what and how to revise the materials.

APPENDIX E

STUDENT BY ITEM MATRIX

STUDENT BY ITEM MATRIX

Attitude Survey
AH 111 SLATE on Swine Breeds

Items

Students	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	Total Score
Remington																															109
Phelps					X															X							X				97
Singer			X		X						X				X												X				93
Dowd		X					X																								111
D'Adamo		X																		X						X	X				93
Taylor		X								X					X												X				100
Welpy		X																		X							X				107
Rosalak																															108
Kart																															112
		4																		3							5				103.33

X = Student response more than 2 points from the extreme.

NOTE: Items 1, 20, and 27 were discussed at the debriefing.

APPENDIX F

BACKGROUND INFORMATION ON THE THREE PARTICIPATING AUTHORS

Table 12.--Background Information on the Three Participating Authors

	Affiliation & Position	Degree & Teaching Experience	Previous Training or Experience in SLATE Design	Instructional System Using the Prototype SLATEs
Author A	MSU Professor	Ph.D. 17 Years Teaching	Had previously developed 10 SLATEs with some professional help from MSU Media Center	3 credit hour course in Animal Husbandry Lecture + lab + 10 SLATEs 175 students per term primarily freshmen & sophomores
Author B	MSU Instructor	M.A. 8 Years Teaching	None	3 credit hour course in Industrial Arts teacher training. Lecture + lab. 40-60 students per term primarily juniors & seniors
Author C	MSU Professor	Ph.D. 22 Years Teaching	Previously developed 8 SLATEs on his own. Did all design & production without professional help. These SLATEs were modeled after the AVT Biology units developed by Postlethwait.	4 credit hour introductory course in Biology. 2 lectures, 1 lab and 1 SLATE per week. 100 students per term primarily freshmen.

APPENDIX G

STUDENT REACTIONNAIRE

STUDENT REACTIONNAIRE

NAME _____ DATE _____

LESSON TITLE _____

Please be frank and honest in answering the following questions. Remember, you are our prime source of information regarding what needs to be revised.

KEY: 1 means you strongly agree; 2 means you agree; 3 means you are uncertain; 4 means you disagree; and 5 means you strongly disagree.

1.	I had sufficient prerequisites to prepare me for this lesson.	1	2	3	4	5
2.	I was often unsure of what, exactly, I was supposed to be learning.	1	2	3	4	5
3.	After completing the lesson, I felt that what I learned was either directly applicable to my major interest, or provided important background concepts to me.	1	2	3	4	5
4.	Manipulating the equipment, or equipment breakdowns often distracted my attention.	1	2	3	4	5
5.	Listening to the tapes and watching the slides became tedious, or boring.	1	2	3	4	5
6.	This lesson was very well organized. The concepts were highly related to each other.	1	2	3	4	5
7.	A professional speaker (announcer) should be used to make the tapes.	1	2	3	4	5
8.	The audio tape moved too fast for me, there was too much information.	1	2	3	4	5
9.	There was too much redundancy. I was bored by the repetition of ideas.	1	2	3	4	5
10.	There was a lot of irrelevant information in this lesson.	1	2	3	4	5

11.	The workbook was excellently designed. I could easily follow the instructions and perform the exercises.	1	2	3	4	5
12.	Frequent reference to and use of the workbook was distracting.	1	2	3	4	5
13.	Often the tape and slides seemed unrelated to each other.	1	2	3	4	5
14.	This lesson had very serious gaps and lacked internal continuity.	1	2	3	4	5
15.	The examples used to illustrate main points were excellent.	1	2	3	4	5
16.	The vocabulary used contained many unfamiliar words. I often did not understand what was going on.	1	2	3	4	5
17.	The pre-test and final exam questions did a good job of testing my knowledge of the main points in the lesson.	1	2	3	4	5
18.	The questions during the lesson gave me valuable feedback on how I was doing.	1	2	3	4	5
19.	Many of the things I was asked to do, or questions I was asked to answer during the lesson seemed like needless busy work.	1	2	3	4	5
20.	At the end of the lesson I was still uncertain about a lot of things and had to guess on many of the final exam questions.	1	2	3	4	5
21.	I believe I learned a lot, considering the time spent on this lesson.	1	2	3	4	5
22.	I would recommend extensive modifications to the lesson before using it with other students.	1	2	3	4	5

23. For you, what was the most difficult part of the lesson? _____

_____

24. What was the easiest part of the lesson? _____

_____

25. What were the three worst things about this lesson? _____

_____

_____

_____

		1	2	3	4	5
26.	I understood most of the concepts and vocabulary immediately after completing the lesson.					
27.	I think this whole procedure of trying out new materials with students is a waste of time.					
28.	I would prefer a textbook or lecture version of this lesson rather than the slide/tape/workbook version.					
29.	I often needed to go back over a portion of the lesson to fully understand it.					
30.	After completing the lesson, I was more interested in and/or favorably impressed with the general subject matter than I was before the lesson.					

31. Please write below any comments, suggestions, or changes which you believe will improve this lesson. Thank you.

APPENDIX H

TRYOUT "CHECKLIST" AND INTERVENTION PRINCIPLES

TRYOUT "CHECKLIST" AND
INTERVENTION PRINCIPLES

The Tutorial Approach

1. The programmer should first explain to the tryout student that the materials he is to be given are intended to help him learn subject matter designated in the title.
2. The programmer should emphasize that the role of the student is to help the programmer evaluate some new education materials. Comments and suggestions that the student makes will help the programmer make revisions.
3. The programmer should then explain that he has to know how much the student already knows about the subject matter and whether or not the student has all of the prerequisites to learn from the materials. He should then give the student the pre-test (always) and the prerequisites test (if required) timing the student on both. Both of these may be done when the test subjects are being selected.
4. When the tests have been completed, the programmer should show the student the program and explain again that it is the material, not the student, that is to be tested from now on. This is an especially important point about which the student should have no question.
5. The student should be given a ball point pen with which to write his answers. (This will prevent him from erasing potentially valuable information for revising the program.) He should be provided with answer sheets, if any.
6. Tell the student to put an "X" next to the items he thinks he got wrong after he has checked his answer. If the program contains open-ended questions, tell the student about this.
7. Explain to the student that if he doesn't know an answer, he should take a guess and write "guess" on the answer sheet. If he simply can't think of an answer, he should leave the answer blank and place an "X" next to the item on the answer sheet.
8. Tell the student the time limits placed on the tryout session and that he can take a break whenever he feels like stopping.
9. Re-emphasize that any comments he wants to write or express to the programmer will be useful and welcomed.
10. Then ask the student to commence with the materials. (If the student asks what he should do or asks if he's doing it right, the programmer should gently insist that all the directions necessary are given in the materials. It is important to try out the directions, too.)
11. The programmer should note carefully the time at the beginning and end of each tryout session and keep track of "break time."

Checklist for the First Tryout Sessions (Horn, 1966, p. 6)

The Tutorial Approach

Principles

I. If the student can continue through the program even though he has difficulty with an item, it is best to let him continue. Ask him about the difficulty at the end of the tryout session. Watch him very carefully for three or four frames. If he's consistently in trouble, it may be well to interrupt.

II. If the student has so much difficulty with an item that he cannot proceed with the rest of the program, the programmer should intervene. His first step should be to try to revise the program on the spot, presenting a revised or new item to the student. This may be done orally or the programmer may make written changes in the program. He should do this revision with a minimum of explanation to the student.

III. If these on-the-spot revisions do not work or if the programmer can't figure out the difficulty, he may then query the student directly with such open-ended questions as: "Will you tell me about the difficulty?" or "What seemed to be the trouble with this item?"

How to Intervene in the Tryout Process (Horn, 1966, p. 12)

APPENDIX I

RULES TO BE FOLLOWED FOR THE REVISION OF A CALCULUS PROGRAM

RULES TO BE FOLLOWED FOR THE
REVISION OF A CALCULUS PROGRAM

The method proposed to the writers for using data for the revision of the programmed materials:

1. Study the item analysis of the end-of-lesson test to determine those concepts which were most often missed by the students.

2. Study the incorrect responses to these particular test items to determine if there was a straightforward misunderstanding of notation, a complete lack of comprehension of the concept, or a variety of errors.

3. Use the guide to determine those frames in the program which dealt most directly with the concept(s) missed on the test.

4. Study the student error rates for these frames. If the program frames are quite similar to the test item, and the error is quite low, more practice frames should be provided. If the error rate is quite high, these frames need revision.

5. Study the sample of incorrect student responses to this segment of the program. These responses should suggest the nature of the learning difficulty and the type of revision needed.

6. Study the comments of both the students and the program reviewers for further suggestions concerning the problems encountered with these particular frames.

7. If no frames in the program correspond to a test item missed by a large percentage of the students, consider the addition of frames that will "bridge the gap" between the present learning materials and what would be considered a transfer type item.

Rules to be Followed in Revising a Calculus Program (Dick, 1968, p. 100)

APPENDIX J

SLATE A_1 RAW DATA

Table 13.--SLATE A_1 Raw Data

CONTROL GROUP (N=12)

Student	Pre-Test	Post-Test	Gain Score	Attitude Survey	(38 correct) 80% Criterion
A	29	41	12	79	Yes
B	23	38	15	107	Yes
C	25	37	12	100	No
D	31	41	10	108	Yes
E	18	32	14	103	No
F	20	37	17	75	No
G	20	39	11	104	Yes
H	19	27	8	71	No
I	31	40	9	98	Yes
J	22	39	17	97	Yes
K	20	39	19	98	Yes
L	22	36	14	102	No
$\bar{X}$	23.33	37.17	13.83	95.17	7 out of 12 for 58.33%

NOTE: Pre- and post-test raw scores based on 46 common items worth 47 points maximum.

Attitude survey raw scores based on 27 item rating scale instrument worth 135 points maximum.

EXPERIMENTAL GROUP (N=12)

Student	Pre-Test	Post-Test	Gain Score	Attitude Survey	(38 correct) 80% Criterion
A	23	45	22	105	Yes
B	14	28	14	91	No
C	27	41	14	95	Yes
D	20	43	23	110	Yes
E	22	41	19	116	Yes
F	18	45	27	99	Yes
G	21	42	21	113	Yes
H	23	44	21	105	Yes
I	26	46	20	114	Yes
J	22	44	22	110	Yes
K	17	46	29	119	Yes
L	22	43	21	102	Yes
$\bar{X}$	21.25	42.33	21.08	106.58	11 out of 12 for 91.6%

APPENDIX K

SLATE A_2 RAW DATA

Table 14.--SLATE A_2 Raw Data

CONTROL GROUP (N=12)

Student	Pre-Test	Post-Test	Gain Score	Attitude Survey	(32 correct) 80% Criterion
A	15	24	9	93	No
B	17	27	10	118	No
C	23	29	6	116	No
D	18	30	12	107	No
E	12	27	15	105	No
F	6	25	19	97	No
G	14	32	18	120	Yes
H	9	34	25	107	Yes
I	26	32	6	110	Yes
J	16	34	18	86	Yes
K	10	23	13	96	No
L	11	35	24	105	Yes
$\bar{X}$	14.75	29.33	14.58	105.0	5 out of 12 for 58.33%

NOTE: Pre- and post-test raw scores based on 40 common items worth 40 points maximum.

Attitude survey raw scores based on 27 item rating scale instrument worth 135 points maximum.

EXPERIMENTAL GROUP (N=9)

Students	Pre-Test	Post-Test	Gain Score	Attitude Survey	(32 correct) 80% Criterion
A	13	30	17	102	No
B	14	30	16	89	No
C	15	34	19	110	Yes
D	17	38	21	109	Yes
E	8	23	15	108	No
F	18	38	20	105	Yes
G	19	35	16	126	Yes
H	17	37	20	102	Yes
I	27	36	9	114	Yes
$\bar{X}$	16.44	33.44	17	106.40	6 out of 9 for 66.6%

APPENDIX L

SLATE A_3 RAW DATA

Table 15.--SLATE A_3 Raw Data

CONTROL GROUP (N=9)

Student	Pre-Test	Post-Test	Gain Score	Attitude Survey	(40 correct) 80% Criterion
A	44	50	6	108	Yes
B	43	47	4	109	Yes
C	22	43	21	93	Yes
D	25	36	11	112	No
E	23	48	25	107	Yes
F	21	40	19	93	Yes
G	19	42	23	111	Yes
H	22	38	16	97	No
I	18	45	27	100	Yes
$\bar{X}$	26.33	42.22	16.89	103.44	7 out of 9 for 77.7%

NOTE: Pre- and post-test raw scores based on 50 items worth 50 points maximum.

Attitudinal survey raw scores based on 27 item rating scale instrument worth 135 points maximum.

No experimental treatment conducted in A_3.

APPENDIX M

SLATE B_1 RAW DATA

Table 16.--SLATE B_1 Raw Data

CONTROL GROUP (N=7)

Student	Pre-Test	Post-Test	Gain Score	Attitude Survey	(12 correct) 80% Criterion
A	0	5	5	78	No
B	2	11	9	102	No
C	10	12	2	97	Yes
D	3	9	6	88	No
E	7	12	5	86	No
F	2	7	5	77	No
G	2	13	11	94	Yes
$\bar{X}$	3.71	9.86	6.14	88.86	42.85%

NOTE: Pre- and post-test raw scores based on 15 common items worth 15 points maximum.

Attitude survey raw scores based on 27 item rating scale instrument worth 135 points maximum.

EXPERIMENTAL GROUP (N=8)

Student	Pre-Test	Post-Test	Gain Score	Attitude Survey	80% Criterion
A	5	13	8	104	Yes
B	6	13	7	123	Yes
C	5	15	10	92	Yes
D	4	15	11	123	Yes
E	5	15	10	126	Yes
F	6	15	9	117	Yes
G	4	14	10	106	Yes
H	5	14	9	105	Yes
$\bar{X}$	5.00	14.25	9.25	112.00	100%

APPENDIX N

SLATE C_1 RAW DATA

Table 17.--SLATE C_1 Raw Data

CONTROL GROUP (N=14)	
Student	Attitude Survey
A	79
B	85
C	87
D	94
E	97
F	99
G	107
H	93
I	98
J	97
K	86
L	100
M	104
N	113
$\bar{X}$	95.64

NOTE: No pre- and post-tests were developed for C_1.

No experimental treatment was developed for C_1.

Attitude survey raw data based on 27 item rating scale worth 135 points maximum.