

IDENTIFICATION OF CRITICAL VARIABLES IN DEVELOPMENTAL TESTING AND AN EXPERIMENTAL
EXAMINATION OF THE NUMBER AND ROLES OF PARTICIPANTS IN TESTING SESSIONS

Charles Wesley Burt

Department of Educational Psychology and Counselling

McGill University, Montreal

April, 1989

A thesis submitted to the Faculty of Graduate Studies and Research

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

ABSTRACT

A large sample ($N = 206$) of expert instructional designers was surveyed in order to determine how developmental testing is designed and conducted. The majority of respondents indicated that they did conduct formative testing of prototype instructional materials for purposes of revising them. Both one-to-one and small group testing were recommended with both the developer and the test subjects assuming active roles during the testing sessions. Both quantitative and qualitative (including attitudinal) data are gathered with an emphasis placed on the analysis of the posttest data. Based on a review of the literature and the results of the survey, an experiment was conducted in order to investigate the effects of different roles that the participants might play during testing sessions with different numbers of participants within sessions. Individuals assigned to small groups did not produce as many comments during instruction as the one-to-one procedure. A content analysis of the qualitative data revealed that test subjects generally comment by indicating comprehension/recall problems. As a source of feedback, the attitudinal survey provided a considerable amount of data especially with small groups which were reluctant to make critical comments during instruction.

RÉSUMÉ

Un sondage a été fait à partir d'un échantillon ($N=206$) de dessinateur instructionnels, experts dans leurs champ, afin de déterminer comment le testing développemental est projeté et conduit. La majorité des répondants ont indiqués qu'ils ont conduits le testing formatif de matériaux prototype instructionnel dans le but de les reviser. À la fois, le testing dans les conditions d'un élève avec un évaluateur et le testing avec des petits groupes et un évaluateur ont été recommandés ayant le évaluateur et les étudiants jouant un rôle actif pendant les sessions de testing. À la fois, les données quantitatives et qualitatives (incluant attitudinal) sont obtenus avec une emphase placée sur l'analyse des données de le posttest. Fondée sur une revue du matériel écrit et de les résultats du sondage, une expérience fut conduit dans le but d'investiguer les effets des différents rôles que les participants pourront jouer pendant les sessions de testing avec un nombre différent de participants à l'intérieur des sessions. Les individus qui ont été assignés à des petits groupes n'ont pas produits autant de commentaires pendant l'instruction comparés aux sessions où il y avait seulement un élève par évaluateur. Une analyse du contenu de l'information qualitative a révélé que les étudiants font généralement des commentaires en indiquant leurs problèmes de compréhension et de rappel. Comme une source de feedback, le sondage attitudinal a fournit un montant considérable de données, surtout avec les petits groupes qui étaient peu disposés à faire des commentaires critiques pendant la session d'instruction.

Acknowledgement of Contributors to the Research

Parts of the research presented in this doctoral dissertation have been presented elsewhere. The following papers have been used extensively in the preparation of this document. Of special note are the contributions made by Dr. George L. Geis and Dr. Cynthia B. Weston. The contributions of these researchers have been recognized by the author of this document.

Burt, C., & Geis, G. (1986, April). Guidelines for developmental testing: Proposed and practiced.

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 270 500)

Burt, C. (1987, April). An experimental study of group size and participants' role in developmental testing. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 292 853)

Geis, G., Burt, C., & Weston, C. (1984, April). Instructional development: Developmental testing. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 243 793)

Weston, C., Burt, C., & Geis, G. (1984, April). Instructional development: Revision procedures. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

ACKNOWLEDGEMENTS

I would like to extend my gratitude to Dr. George Geis, my committee chairman, for his continuous guidance and support during the conception, elaboration, and production of this research. His collegueship and friendship were instrumental towards the development of my confidence as a professional. Of special note are the similar contributions by Dr. Cynthia Weston. Significant contributions provided by the other members of my committee, Dr. Harold Stolovitch and Dr. Patricia Cranton are also appreciated. Of equal note were the contributions made by all of the graduate student members of the research team.

I would like to extend my thanks to the instructional designers who took the time to complete the questionnaire. Their responses encouraged me to further persue the research which was undertaken. Special thanks are extended to the students who participated in the study. Hopefully students like them, in the future, will benefit from the time and effort which the nursing students provided in order that data could be collected.

Never to be forgotten, was the continual support of my parents, Otto and Peggy, as they saw my dreams become reality. Without their support, in so many different ways, this dissertation could never have been. I profoundly extend my love to them for having encouraged me through many of the difficulties in life. There is not a doubt in my mind, that they provided the foundation upon which all my endeavors have been actualized. This dissertation is dedicated to them.

Many thanks are extended to Dr. Socrates Rapagna who was always there with much needed advice on the statistical analysis of the data. To all of my teachers at McGill, I would like thank them for having provided, for me, the conditions of learning. Friends made at McGill and in Montreal are not to be forgotten either.

This research and its dissemination have been funded by an FCAR grant and support from McGill University.

LIST OF TABLES

	Page
1. Categorization of Guidelines Proposed by Selected Authors for Designing and Conducting Developmental Testing	79
2. Questionnaire Respondents and Non-Respondents	88
3. Demographic Breakdown of Samples' Principal Type of Employment	89
4. Demographic Breakdown of Samples' Most Significant Employment Responsibility ...	90
5. The Mean Number of Test Subjects and Sessions	106
6. Distribution of Samples' Mother Tongue	128
7. Distribution of Samples' Age	128
8. Distribution of Samples' Educational Background	129
9. Distribution of How Some of the Sample Acquired Prior Knowledge	129
10. Scoring Key by Individual Self-Test Item	148
24. Two-Way ANOVA Summary Table of the Number of Problems Identified (Comments) during Developmental Testing	164
38. Two-Way ANOVA Summary Table of the Number of Comments that were Debriefed ...	167
52. Two-Way ANOVA Summary Table of Debriefing Agenda Items Based on Posttest Errors	171
66. Two-Way ANOVA Summary Table of Debriefing Agenda Items Based on the Attitudinal Survey Responses	175
71. Two-Way ANOVA Summary Table of the Total Number of Debriefing Agenda Items from all Sources Combined	178
72. General Categories and Sub-categories for Problem Explication Responses	189
73. General Categories and Sub-categories for Revision Responses	191
76. 2 x 2 MANOVA Summary Table of Categorized Problem Explications for Unit #1	196
77. 2 x 2 MANOVA Summary Table of Categorized Revisions for Unit #1	196
80. 2 x 2 MANOVA Summary Table of Categorized Problem Explications for Unit #2	197
81. 2 x 2 MANOVA Summary Table of Categorized Revisions for Unit #2	197

82. Two-Way ANOVA Summary Table of the Amount by Category of General Kinds of Problems Explicated during the Developmental Testing of Unit #2	201
85. 2 x 2 MANOVA Summary Table of the Total Number of Categorized Problem Explications	203
86. 2 x 2 MANOVA Summary Table of the Total Number of Categorized Revisions ..	203
89. 2 x 2 x 2 MANOVA Summary Table of Categorized Problem Explications for Unit #1 ..	205
90. 2 x 2 x 2 MANOVA Summary Table of Categorized Revisions for Unit #1	205
93. 2 x 2 x 2 MANOVA Summary Table of Categorized Problem Explications for Unit #2 ..	207
94. 2 x 2 x 2 MANOVA Summary Table of Categorized Revisions for Unit #2	207
97. 2 x 2 x 2 MANOVA Summary Table of the Total Number of Categorized Problem Explications	208
98. 2 x 2 x 2 MANOVA Summary Table of the Total Number of Categorized Revisions	208

LIST OF FIGURES

	Page
1. A General Model for Human Performance	4
2. A Research Programme Investigating Developmental Testing Procedures	21
3. Role of Test Subject	34
4. Role of Developer/Evaluator	36
5. The Bases of the Revision Process	41
6. Procedures Used for the Design and Conduct of Developmental testing	102
7. The Procedures Used Within the Experimental Design	131
8. 2 x 2 MANOVA Data Analysis Design #1	133
9. 2 x 2 x 2 MANOVA Data Analysis Design #2	134
10. Plotting of Means: Main Effects of Role and Number of Participants on the Number of Problems Identified by Test Subjects Commenting during Developmental Testing ..	165
11. Plotting of Means: Interaction between Role and Number of Participants with the Number of Problems Identified by Test Subjects Commenting during Developmental Testing	166
12. Plotting of Means: Main Effects of Role and Number of Participants on the Number of Debriefing Agenda Items Based on Comments	168
13. Plotting of Means: Interaction between Role and Number of Participants with the Number of Debriefing Agenda Items Based on Comments	169
14. Plotting of Means: Main Effects of Role and Number of Participants on the Number of Debriefing Agenda Items Based on Posttest Errors	172
15. Plotting of Means: Interaction between Role and Number of Participants with the Number of Debriefing Agenda Items Based on Posttest Errors	173
16. Plotting of Means: Main Effects of Role and Number of Participants on the Number of Debriefing Agenda Items Based on Survey Responses	176
17. Plotting of Means: interaction between Role and Number of Participants with the Number of Debriefing Agenda Items Based on Survey Responses	177
18. Plotting of Means: Main Effects of Role and Number of Participants and their Interactions with the Total Number of Agenda Items (from all feedback sources)	

Debriefed.....	180
21a. Plotting of Means: Interactions of Role and Number of Participants with Categorized Problem Explications for Unit #2.....	198
21b. Plotting of Means: Interactions of Role and Number of Participants with Categorized Problem Explications for Unit #2.....	198

TABLE OF CONTENTS

	Page
RESEARCH ACKNOWLEDGEMENT	I
ACKNOWLEDGEMENTS	II
LIST OF TABLES	III
LIST OF FIGURES	IV
.....	
Chapter	
I. INTRODUCTION	1
Context of the Study	
Role of Formative Evaluation Within a Model for Instructional Design	
Formative Evaluation and Common Variants:	
- Formative Evaluation	
- Learner Verification and Revision	
- Three-stage Model	
- Developmental Testing	
- Expert/Internal Review	
A Comparison of Formative Evaluation Procedures Within a Chronology of Phases	
Learner Input During the Instructional Design Process	
Description of the Problem and Purpose of the Study	
Research Questions	
II. REVIEW OF THE LITERATURE	25
Instructional Systems Development Models	
Isolation of Developmental Testing Variables	
Instructional Design Textbook Literature Review:	
- The Criteria for Acceptable Student Performance	
- The Number of Test Subjects (Learners), the Number of Sessions Required During Testing, and General Testing Procedures	
- The Characteristics of the Test Subject Population	
- The Role of the Test Subject(s) During Developmental Testing and Revision	
- The Role of the Designer/Evaluator During Developmental Testing and Revision	
- The Method of Collecting Data	
- The Basis for Revision	
- The Testing Environment	
A Summary of the Textbook Literature of Guidelines	
Critical Analysis of Empirical Studies:	

- Internal Review
- Field Testing
- Learner Verification and Revision (LVR)
- Developmental Testing

Comparisons of Approaches
 Generalizations Based on the Literature review
 Research Questions

III. QUESTIONNAIRE ADMINISTRATION 77

Rationale for the Questionnaire
 Designing the Questionnaire
 Pilot Testing
 Sampling Procedures
 Mailing Procedures
 Post Hoc Statistical Hypotheses Generation
 Data Analysis Procedures
 Results

IV. A COMPARISON OF RECOMMENDED AND ACTUAL DEVELOPMENTAL TESTING PRACTICES 103

Discrepancies Between Recommended and Actual Practices
 A Profile of Expert Procedures
 Discussion
 Conclusions
 Research Questions

V. AN EXPERIMENTAL MANIPULATION OF DEVELOPMENTAL TESTING VARIABLES . . . 120

Rationale for the Experiment
 Description of the Instructional Materials:

- Criteria Used to Select the Stimulus Materials Employed in the Experiment
- A History of the Development of the Materials
- Differences Between Unit One and Unit Two of the Materials

Description of the Population Utilized and Sampling Procedures
 Experimental Design and Research Questions
 Study Procedures:

- All conditions
- One-to-One Active
- One-to-One Passive
- Small Group Active
- Small Group Passive

Instrumentation and Data Sources
 Criteria Used to Construct the Debriefing Agenda
 Different Debriefing Strategies Utilized Between One-to-One and Small Group
 Results of Pilot Testing the Experimental Procedures
 Coding the Amount of Data Obtained from the Different Sources of Feedback:

- Source: Comments Made During Instruction
- Source: Debriefed Comments
- Source: Debriefed Self-Test Item Errors
- Source: Debriefed Attitudinal Survey Results
- The Total Amount of Feedback Obtained

Relationships Between the Various Sources of Feedback Data:

- Relating Comments Made During Instruction to Comments About the Instruction Made During Debriefing
- Relating Self-Test Performance to Comments Made During Instruction
- Relating Self-Test Results to Test Subject's Perceptions of Self-Test Error Causation
- Relating Comments Made While Taking the Self-Test to Comments Made After Knowledge of Results of the Self-Test
- Relating Comments Based on Knowledge of Attitudinal Survey Results with any Previous Feedback

Different Kinds of Test Subject Comments

Content Analysis of Comments: Explicated Problem and Revision Responses

VI. DATA ANALYSIS PROCEDURES AND RESULTS OF THE EXPERIMENTAL MANIPULATION OF DEVELOPMENTAL TESTING VARIABLES 161

Data Analysis Procedures and Results by Research Question

VII. DISCUSSION OF THE RESULTS OF THE EXPERIMENT. 209

VIII CONCLUSIONS 226

IX HEURISTICS FOR DESIGNING AND CONDUCTING DEVELOPMENTAL TESTING SESSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH 234

Heuristics for Developmental Testing

- Selecting the Test Subjects
- One-to-One Testing and Debriefing
- Small Group Testing

Recommendations for Future Research

- Preparing Subjects
- Other Media
- Identifying Responders
- Revising Instruction
- On-line versus Retrospective Data Collection
- Embedded Problems

REFERENCES **242**

APPENDICES **251**

APPENDIX A

APPENDIX B

APPENDIX C

APPENDIX D

APPENDIX E

APPENDIX F

APPENDIX G

APPENDIX H

APPENDIX I

APPENDIX J

APPENDIX K

CHAPTER I

INTRODUCTION

Context of the Study

There is an expressed interest by those professionals who design instructional materials according to systematic instructional design principles that instruction should reflect an iterative cycle of evaluation for purposes of improving the instruction. Many refer to this quality control aspect of instructional design as formative evaluation. Recently, the Journal of Instructional Development (Spring, 1980), Performance and Instruction Journal (June, 1983), and the Canadian Journal of Educational Communications (Winter, 1986) have devoted full issues of their journals to the topic of formative evaluation. Articles within these publications reflect a commitment to instructional improvement through formative evaluation procedures. This commitment is not novel within education. Tyler (1947) viewed evaluation, even before the term formative evaluation evolved, as a necessary and iterative process through which educational programs could be improved:

The process of evaluation is an integral part of the education process. It does not mean simply the giving of a few ready-made tests and the tabulations of resulting scores. It is a recurring process involving the formulation of objectives, their clearer definition, plans to study students' reactions in the light of these objectives, continued efforts to interpret the results of such appraisals in terms which throw helpful light on the education program and on the individual student. This sort of procedure goes on as a continuing cycle. Studying the results of evaluation often leads to a reformulation and some improvement in the conception of the objectives to be obtained. The results of evaluation and any reformulations of objectives will suggest desirable modifications in the teaching and in the educational program itself. Modifications in the objectives and in

the educational program will result in corresponding modifications in the plan and program of evaluation. So the cycle goes on. (pp. 500 - 501)

Generally, Scriven (1967) is given credit for having distinguished between formative and summative evaluation when he originally coined the terms. Formative evaluation was seen as a mechanism of collecting information that would facilitate making judgements as to how to revise and improve instructional programs. The process requires gathering feedback as to the effectiveness of instruction while it is being developed as opposed to after it has been developed as is the case in summative evaluation.

The instructional design literature offers numerous models or heuristics for the design of replicable instruction. It makes little sense, in any cost effective way, to apply elaborate formative evaluation procedures to instruction that cannot be accurately described and repeatedly used by learners. By applying instructional design heuristics, teaching systems which are both accurate in content and effective in producing learning can be engineered (c.f., Davis, Alexander, & Yelon 1974). The cost and effort involved in carrying out the complete process has led to applications which result in the development of reusable instructional materials rather than in less permanent classroom teaching. Materials produced following the prescribed heuristics should be efficient means for individualizing instruction (i.e., matching learner needs to the appropriate materials). Often these materials are designed to be self-instructional modules that a learner can use on a self-paced basis.

Formative evaluation is one component of the design process that has been consistently recommended by most design models. Andrews and Goodson (1980) reviewed 40 instructional design models and found that 38 included formative evaluation as a major component. Thus a large majority of models do recommend that instructional materials be empirically tested using test subjects (students/learners) in order to obtain feedback data as to the effectiveness of the materials in achieving specified learning objectives. Stolovitch (1982) reached a similar conclusion after having systematically reviewed 12 internationally employed instructional development systems models. The use of students as test subjects was found to be an activity prescribed in 11 out of the 12 reviewed

models. Gropper (1975) and Nathenson and Henderson (1979) have recognized the importance of using student feedback to improve learning materials and have written textbooks specifically on the topic. Although there is strong endorsement to evaluate instruction as it is being designed, there is little consensus across models in specifying which feedback/revision procedures to utilize and little evidence to indicate the relative effectiveness of the varying procedures being proposed in the literature.

Role of Formative Evaluation within a Model for Instructional Design

Figure 1 presents a model for instructional design, adapted from the work of Geis and Weston (1983), that is not unlike those found in the literature (c.f., Gagne and Briggs 1979; Reigeluth 1983; Dick and Carey 1985; and Knirk and Gustafson 1986) and in that respect can be considered prototypic of the plethora of existing models. For example, Merrill and Boutwell (1973) after having reviewed textbooks and manuals which were specifically designed to provide guidance to instructional systems developers, noted that these sources varied their terminology but basically prescribed similar procedures in the form of the following components: behavioral objectives, pretest, instructional activities, posttest, and revision based on empirical results. Formative evaluation is geared to improving instruction through the revision process.

In their reviews of formative evaluation procedures of instructional materials, Baker (1974), Senders and Cunningham (1973; 1974), Thiagarajan (1976), Kandaswamy (1980), Cambre (1981), and Weston (1986), all discuss different variations that have come to be associated with the literature of formative evaluation. Generally, formative evaluation and related variants use some combination of the following phases:

1. by the original designer reviewing the instruction — often referred to as self-evaluation
2. by an external expert reviewing the instruction or a variety of different kinds of experts acting as a jury for review purposes (e.g., content experts, test specialists, media consultants,

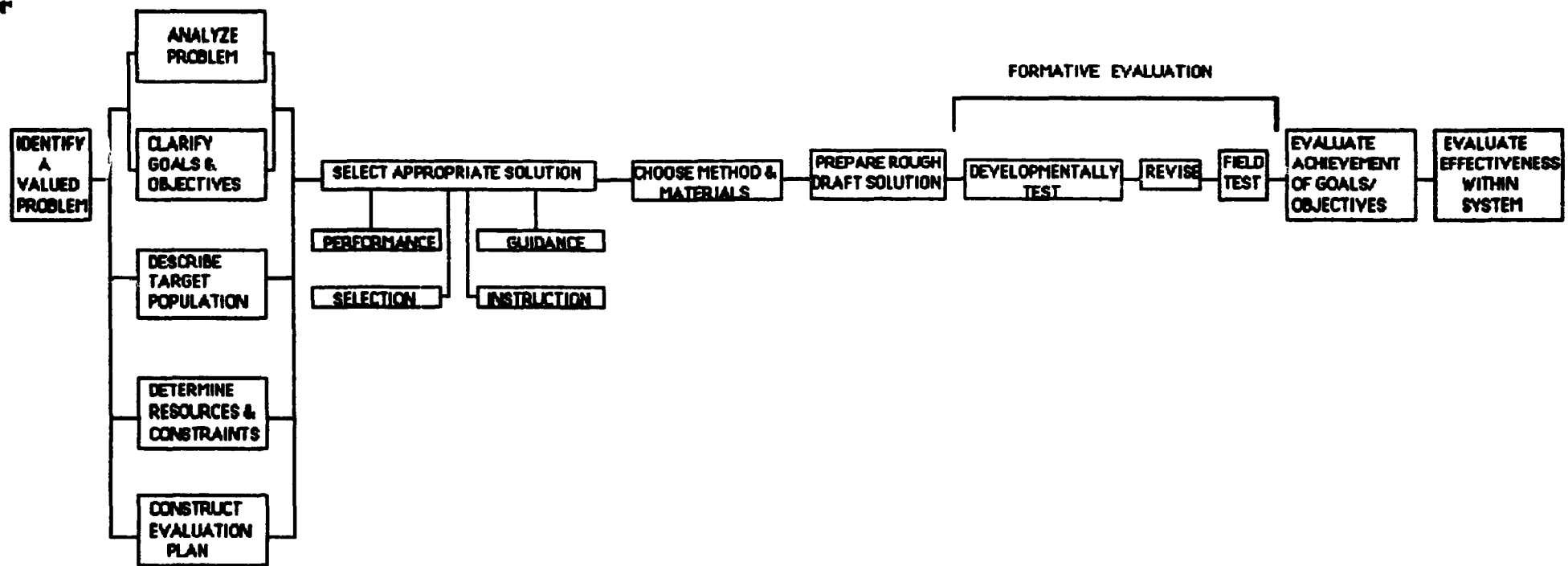


Figure 1. A general model for human performance (adapted from Geis & Weston, 1983)

etc.)

3. by a series of one-to-one clinical trials of the instruction with the designer/evaluator sitting with individual student test subjects as they go through the instructional materials
4. by trying out the instruction with a small group or groups of test subjects
5. by field testing the instruction with a larger group or groups of students in order to approximate the way in which instruction would ultimately be delivered
6. by extending the testing of the instruction for as long as it continues to be delivered to learners (usually included in learner verification and revision models — LVR)

There are several reasons for evaluating instruction as it is being developed. During formative evaluation, errors that learners make and negative reactions to the instruction at this stage of the design process are assumed to be largely the result of faulty design unless there is evidence to the contrary. Because the literature of instructional design offers heuristics as opposed to algorithms for design, instructional decisions and possible assumptions made while following a design heuristic could have been faulty. Smith and Murray (1975) recognize the heuristic nature of design models by stating that:

Development models, however, frequently appear to be as much managerial strategies as helpful conceptual paradigms of any particular development process. The characterization of models as heuristic devices for use in specific contexts suggests a conceptual difficulty in generalizing or being prescriptive about the use of particular models across varying circumstances and situations. (p. 13)

Much of the literature on instructional design is directed towards practitioners and many authors view this literature as being based more on practice than on theory due to weaknesses in existing theories. Knirk and Gustafson (1986) cite weaknesses in instructional theory as the main reason for the need to try out instruction before it is released:

... theories of instruction are needed as part of the research process in organizing information about human learning. However, there is still a long way to go before any of these theories will adequately prescribe a process for consistently designing effective instruction for individuals or groups of students. At the present time they serve principally as useful guides or frameworks within which instructional design can take place. Due to this limitation of current theories, the design process places great emphasis on testing a proto-type of the instruction before disseminating it for general use. (p. 105)

Similar criticisms have been levied against learning theory to the point where Markle (1967) concluded that:

Despite all the laboratory research in learning in many years of psychological science, no formula exists for producing the right amount or schedule of practice to establish a particular skill or for determining the right number or range of examples to establish a particular generalization. These are, in a clear sense, factors of that variable called 'size of step' and are ways in which the strategy for one student may differ from the optimum strategy for another. (p. 122)

Baker (1973) implies that it could be limitations of both learning and instructional theory that mandate the need for the testing of instruction during the prototype stage of developing instruction.

The acceptance of prototype development as a regular procedure in design is a clear admission of the inadequate knowledge base and state of the instruction art. It is a recognition that there is but limited understanding of how learning can best be facilitated and that present information is in no way generalizable within all learners or

to products which have not yet been developed. Thus, prototype design inherently requires empirical data to validate preliminary decisions. (p. 262-263)

Formative evaluation of the prototype instruction, based on empirical tryout data may reveal, for example, that the needs assessment was faulty, the task analysis was not valid, some of the test items were ambiguous, all of the identified pre-requisites were not required, a film would be a better medium of presentation, or that more practice was needed in order to achieve transfer. Errors made during design or misconceptions brought to a formative evaluation should be revealed by the process of formative evaluation. Appropriate revisions are made in light of the data. This quality control aspect of evaluation emphasizes the engineering approach to designing and delivering instruction. It also insures that an individual learner's needs will be met at the point where instruction is failing to meet those needs. At that point, instructional alternatives can be explored and evaluated with the intent that learning will be facilitated.

The last phase of instructional design is to summatively evaluate the effectiveness of the instruction. Usually larger numbers of students are used to test the instruction within a setting that simulates the normal conditions under which the instruction would be typically delivered. Achievement test results are often the prime indication of the success or failure of student learning within the particular curriculum design.

Formative Evaluation and Common Variants

Weston (1986; 1987a) presents a framework for overviewing the different ways that formative evaluation has been commonly conceived and approached in the past. She identifies five variants of formative evaluation that appear frequently in the literature of instructional design. They include: 1) formative evaluation 2) learner verification and revision 3) a 3-stage model 4) developmental testing and 5) expert/internal review. Her synthesis of the formative evaluation literature related to improving instructional materials expands on the work of Sanders and Cunningham (1973), Baker

(1974), Passmore, Asche, and Chase (1975), Thiagarajan (1976), and Cambre (1981). All of these authors, at different times, have formulated broad classification schemes for summarizing the relationship between the different sources of both quantitative and qualitative data applicable to the different phases of formative evaluation. What emerges is a continuum of phases of evaluation activities with the results of each phase providing input for the next phase.

Formative Evaluation

The first variant, formative evaluation as a broad and generic activity, can be seen in its most comprehensive form as a series of procedures that begins with the original author of the instructional materials going back to self-critique and revising the work in its draft form. This is referred to as self-evaluation. Expert review, or internal review as it is commonly referred to, is then utilized to point to content inaccuracies or omissions, problems with the sequencing of topics, and any other qualitative concerns that different kinds of experts might identify. Following this expert/internal review procedure is one-to-one testing where a student test subject, belonging to the target population for which the instruction was produced, is asked to go through the revised materials with an evaluator or often the designer, try to learn from them, do any exercises contained within the instructional package, and complete a posttest. Problems encountered by the test subject can be addressed and revisions made to any part of the instruction that failed to support the test subject in attaining the objectives. After several one-to-one sessions the instruction is given to a small group or to groups of test subjects, which more closely approximates a classroom situation, in order to identify patterns of problems or errors that may not have surfaced during one-to-one testing — possibly due to the limited number of test subjects utilized during one-to-one testing. After revisions are made based on the group data, the instructional materials are field tested under conditions that should closely approximate those that will be encountered by the target population. The developmental state of the materials is somewhat different during field testing than in any other phase utilizing student test subjects. Field testing usually involves a complete instructional package whereas any of the preceding phases tend to evaluate

smaller portions or individual sections of the rough draft materials. Field test data can be collected from teachers, students, or anyone else who will be associated with the implementation of the instruction. Larger groups of students, e.g., intact classrooms, are often utilized. This description of formative evaluation has included all of the possible stages that could be used within a formative evaluation and in this respect can be considered as the broadest definition of the term. However, as Cambre (1981) and Weston (1986) have both pointed out, formative evaluation has been generally conceived of as any of the above procedures, used separately or in some combination in order to review and revise the instruction.

Learner Verification and Revision

A second variant of formative evaluation is Learner Verification and Revision (LVR) which focuses on one-to-one testing, group testing, field testing, and the extended testing or long-term monitoring over the life of the instructional product. The EPIE Institute (1975), Stolovitch (1982), Komoski and Woodward (1985), and the State of Florida's Department of Education (1986) have all expressed concern for the effectiveness of materials that are accepted into instructional systems and will be used over a long period of time, recognizing that both target populations and instructional content may change over time. By continuously monitoring the effectiveness of the instructional material, it is possible to revise the instruction at the point at which it does not serve the target population's needs. Thiagarajan (1978) provides a rationale for extended testing of the instructional product: "The structure of the content, the characteristics of the target population, and the needs of the society change so frequently that continual LVR is required to update an instructional product." (p. 5)

Three-Stage Model

Another variant of formative evaluation is referred to as the three-stage model. This model sequentially incorporates one-to-one testing, small group testing, and field testing. Some of the proponents of this approach include Briggs (1970); Friesen (1973); Thiagarajan, Semmel, and Semmel (1974); Dick (1977); and Dick and Carey (1978; 1985).

Developmental Testing

A fourth identified variant is developmental testing. Some authors (e.g., Lawson 1974; Yelon 1974; Merrill & Tenmyson 1977; and Geis, Burt, & Weston 1984) emphasize utilizing both one-to-one testing and small group testing because both of these conditions allow for the close monitoring of the effectiveness of the instruction by way of the interaction between the designer and the test subjects as they go through the instruction. Developmental testing is interpreted in the literature as either one-to-one testing (e.g., Horn 1966; Bjerstedt 1972; Deterline & Lenn 1972; and Johnson & Johnson 1975), or small group testing (e.g., Abedor 1971; Bell & Abedor 1977).

Expert/Internal Review

A final variant identified within the formative evaluation literature applicable to instructional materials design is expert/internal review. This approach may or may not utilize student test subjects and can rely solely on the expert judgement of appropriate specialists. Stolovitch (1982) makes a distinction between the roles that different kinds of experts can play during internal review. He emphasizes that experts can apply their skills and knowledge only to the particular components of the instruction relevant to areas of expertise in which the expert is competent. These various kinds of experts, e.g., subject matter, technical, legal, etc., do not predict learning outcomes; they consider

local conditions in relation to the objectives resulting in the modification of the instruction prior to going to student testing. Stolovitch (1982) notes that:

Expert appraisal, in the light of local conditions and logically specified objectives, yields a number of modifications which are necessary prior to student testing. Such modifications generally take the form of:

- Corrections in content, language or examples
- Replacement of various visuals considered unsuitable
- Translation to local language
- Recasting of material into locally usable formats (e.g., filmstrip remounted in slide form)

The end result of this phase is a material ready for testing with learners. (p. 20)

On the other hand, diagnostic instruments have been developed for the purpose of aiding an instructional designer in the prediction of specified learning outcomes for a given instructional design (c.f., Merrill, Reigeluth, & Faust 1979 and Montague, Ellis, and Wulfeck 1983). These groups of instructional researchers have conducted studies on the Instructional Quality Inventory (IQI) which is analogous to a checklisting tool that requires the developers to recycle back through all of the components of the design process in order to look for inadequacies and misalignments of design decisions and implementations. This approach considers that review by instructional design experts by itself is sufficient.

In noting the differences among the variants, Weston (1986) notes: "each approach includes a unique combination of evaluation phases; a combination which for each defines the recommended duration of the try-out/revision cycle, the primary source of data for revision and the developmental state of materials during testing." (p. 9)

A Comparison of Formative Evaluation Procedures within a Chronology of Phases

In reference to a chronological sequence of formative evaluation components Kandaswamy (1980) cites Thiagarajan's (1976) model which comprises the following phases:

1. Self-evaluation
2. Professional jurying
3. Individual testing
4. Group testing
5. Field testing
6. Extended testing

Although the terminology, between authors discussing these various activities, varies somewhat, this sequence of activities depicting a chronology of formative phases has logical appeal. Kandaswamy (1980) utilizes a continuum to rationalize this particular sequencing:

... it may be seen that the suggested sequential procedure proceeds along these continua:

1. A shift from formative to summative use of data.
2. A shift from expert appraisal to student performance as the major data source.
3. A shift from the primary goal-based process to a more open goal free process.
4. A shift from developer involvement to third-party evaluation.
5. A shift from soft to hard data.
6. A shift from individual to qualitative feedback.
7. A shift from unitary to multiple instrumentation.
8. A shift from the producer to the consumer (p. 25).

It appears that a number of evaluation phases are recommended because of inherent weaknesses within each particular phase. During the self-evaluation of one's own work, it may be difficult to spot

problems or weaknesses as a result of being "too close" to the work thus adversely affecting the author's ability to be objective. By passing the designed instruction on to an expert or a group of experts, commonly referred to as internal review, the problem of subjectivity that is attached to self-evaluation is reduced. The context in which instructional design efforts are often applied points to the need for a review by other kinds of experts. Instructional designers are usually asked to design instruction in content areas in which they have little expertise. As Bratton (1981) notes, typically a designer initially works with a subject matter expert in order to identify the to-be-learned content. This is one good reason to have the instruction reviewed by a content expert after it has been designed. Geis (1973) and Thiagarajan (1978) note that if instruction is to be declared as valid then questions of content accuracy and relevance need to be addressed. Markle (1983) echoes this concern by stating that:

Expert appraisal and valuing are equally important as student data. Content quality, both input and potential output, is a different issue from learnability. No amount of data from students can validate the quality of potential outcomes or curricula excellence. (p. 29)

A widely endorsed approach for the evaluation of instruction recommends combining the use of appropriate experts (e.g., measurement/evaluation experts, subject matter experts, media/communications experts, developmental psychologists) as consultants within their particular domains of expertise in conjunction with empirical tryout data from members of the target population for whom the material was designed (i.e., student test subjects). This approach, for example, recognizes that students are unable to comment on the validity of the instructional content, but that they might be able to provide useful feedback about the problems they are having while trying to learn from the instruction and they can report on their affective reactions to the learning experience. The combined approach of expert review and collecting tryout data also recognizes, through the research of Newman (1957); Rothkopf (1963); Rosen (1968); Frase, DeGracie & Poston (1974); Henderson

and Methenson (1977); and Smith (1978) that experts are not always successful in predicting what kinds of problems will be encountered by learners (i.e., predicting the effectiveness of the instruction). Within this combined approach, six distinct variations of collecting empirical tryout data from student test subjects have been identified within the instructional design literature:

1. One-to-One
2. Small group
3. One-to-One plus small group
4. Field testing
5. Three-stage model (one-to-one plus small group plus field testing)
6. Learner verification and revision (LVR)

This study focuses on the variables of formative evaluation that involve students as test subjects in a situation that can be closely monitored by an instructional designer/evaluator in order to obtain learner feedback as to the effectiveness of the instruction. This necessarily limits the study by focusing on one-to-one testing and small group testing. Field testing has been ruled out of this particular study because working with large groups, e.g., 30 students or intact classrooms at the site at which the materials will finally be used does not permit a micro-level analysis of the instruction by an instructional designer, e.g., bits of text, several diagrams, or a few slides. Field testing is usually applied to materials that have been fully developed to the point where they are beyond the rough draft stage of development. Field testing assumes that major revisions have already been carried out based on developmental testing and it can be considered as a final validating procedure. Baker (1974), in a review of the different phases of formative evaluation, notes that:

Field tests, by their very nature, are activities that usually involve many people.

Many different classrooms are used, and sometimes different schools, or even school districts, are sampled . . . A program designed for classroom use, with a teacher guiding students through exercises and simulations, should be tried out exactly in such a setting. Only under those conditions can the formative evaluators determine whether the 'great ideas' developed and honed during prototype have any chance of success in the world at large. (p. 562)

One-to-one and small group testing is typically applied to small sections of rough draft materials (prototypes) as they are being developed — not after a whole package has been developed as is the case with field testing. The optimal state of materials development at which formative evaluation is conducted has been addressed in the literature. Research by Rodriguez and Rodriguez (1982) indicates that when formative evaluation is applied to audio-visual learning materials at three different levels of product development (story-board, unprofessional videotape, and professional 16mm film) there are no significant differences in the feedback that students provide to the developer/evaluator. They recommend that if tryout is to be done only once during the production, it should be conducted very early in the development of the product as feedback will be effective at this stage. Modifications done at the early stage will cost less than at later stages of development. Both one-to-one and small group testing are commonly utilized within all of the following variants of evaluation: formative evaluation as a generic activity, Learner Verification and Revision, the three-stage model, and developmental testing.

The term developmental testing originated during the early days of the programmed instruction movement. Then it referred to the close analysis of programmed instructional units consisting of "frames" which tended to consist of a short explanation or a diagram. After each frame, learners were required to make a discrete, overt response based on the instruction that they had just received. An instructional programmer designing auto-instruction would use a typical learner for whom the program was being designed to go through a unit of instructional frames, which was only part of the whole program. By walking, step by step or frame by frame, through the unit with a learner the programmer could track the student's progress and problems on these small sections of instruction. It was the way in which programmed instruction was presented to learners, frame by frame, that encouraged the close analysis of the small bits of instruction. Once a unit was pronounced effective based on try out data, the next unit was developed and then developmentally tested. It was through this process that the program as a whole was validated. Because developmental testing, as a term, has historical precedent based on the literature of programmed instruction (c.f., Horn, 1966) and because of the important contribution that research on programmed instruction has made towards the

understanding of human learning, it will be consistently used to refer to the try out phase of instructional design that incorporates test subject (student) input into determining the effectiveness of instructional materials by way of one-to-one and/or small group evaluation. This study views developmental testing as a procedure of trying out instructional units (often self-instructional) in draft form with a small number of subjects (learners or performers) in order to obtain information to be used in revising and improving the instruction. Revision is seen as a diagnostic and prescriptive process, based on developmental testing, in order to produce instructional materials that meet set performance criteria. Developmental testing and revision are done "in-house" before the learning materials are in final form and ready to be released to the general learner population. As terms, formative evaluation, LVR, and the three-stage model are broader in focus than developmental testing and for that reason were not selected to describe the fine grained, diagnostic function of trying out small units of instructional materials before they are published as a complete package. It should be noted that the developmental testing of instruction delivered by instructors, regular textbooks, films, filmstrips, videotape, interactive video, or computer assisted instruction is not less desirable than the testing of self-instructional print-based materials. It has been recognized by Vedros (1986) that there have been problems in both adapting the process to evaluating textbooks and in getting textbook publishers to adopt the practice even when it is required by law to do so. Just because it might be more difficult to test other mediated instructional delivery systems does not suggest or justify that it is less desirable to do so.

Learner Input During the Instructional Design Process

Developmental testing is only one of the components of the design process that allows for learner input. At the very start, during the initial needs assessment and problem identification phase, students can be asked for both their opinions with regards to the identification of deficits within the current curriculum, and also for their opinions as to how to remedy those deficits. It has also been proposed by Geis (1986) that learners can be effective, integral design partners with the instructional designer in

the planning of instruction. There is research to support such proposals. Gallagher (1984) experimentally investigated the effects of allowing a class of adult learners, master's level educational research students, to make the major design decisions of: sequencing the objectives, defining the target population, determining the instructional strategy, developing and selecting the instruction, and determining the procedures for constructing the criterion-referenced tests. Three classes of students were used in her research design, one which helped design the unit on inferential statistics, another which did not participate in the design process but received the instruction that was cooperatively designed by the other group of students and the instructor, and another group which received strictly instructor designed instruction. No significant differences were found among the three groups on an achievement posttest which suggests that at least adult learners may be able to make equally effective instructional design decisions (i.e., contribute valid input into course planning) as an instructor can. The course planning decisions made within this experiment came at the initial phase of design which is substantially earlier than the developmental testing phase. While it is recognized that learners may be able to provide valid input into many of the phases of instructional design and delivery, the focus of this study will be on designer produced self-instructional materials (instructional modules) with learners providing input during the developmental phase of producing instruction.

Description of the Problem and Purpose of the Study

Systematic instructional design models are based on theory, research outcomes, and practical concerns all of which indicate the effectiveness of instruction produced by using these models. However, the literature abounds with these models as is revealed by Andrews and Goodson's (1980) comparison of 40 different instructional design models which was intended to help a designer in choosing from amongst them. As their review indicates, there is a great deal that is common to all models. Most offer step-by-step prescriptions for creating instructional/training materials. Developmental testing is one task that most models recommend but there is a large amount of variability between models in recommending specific procedures for conducting developmental testing

sessions. The extent of this variability has been reported on by Geis, Burt, and Weston (1984). With such diversity in the literature of instructional design, it is difficult to select, from amongst the numerous suggestions, those guidelines that will enable an instructional designer to get the kinds of data that will be effective for the purposes of revising and improving the instructional design.

In large part, instructional design principles are incorporated into textbooks that provide the theoretical and research based rationales for utilizing specific design procedures. For this study, over 30 different instructional design textbooks and several doctoral dissertations were reviewed. From this pool, those sources that devoted substantial discussion to developmental testing and revision were selected. The substantive selection criteria were based on any given text's provision of enough information to enable a reader to actually conduct developmental testing on a rough draft instructional design. This selection procedure resulted in identifying 14 textbooks and one dissertation which offered specific but at times incomplete guidelines that would frequently contradict with other texts' recommended guidelines. Authors of instructional design textbooks such as Horn (1966), Bjerstedt (1972), and Deterline and Lenn (1972) have made arguments for one-to-one testing using clinical probing of a test subject by the designer as soon as the test subject encounters difficulty. Yet Abedor (1971) and Bell and Abedor (1977) have made a case for developmental testing using a small group of test subjects that try to learn from the material as best they can without being probed by an instructional designer when they encounter difficulties. Authors such as Friesen (1973), Lawson (1974), Thiagarajan et al. (1974), Merrill and Tennyson (1977), and Dick and Carey (1978) recommend using both one-to-one and small group data for revision purposes. It is quite feasible in supposing that the data produced under one-to-one and small group conditions will differ , otherwise why bother conducting product tryouts using both strategies? The literature appears to be divided as to how to collect the data. Conflicting opinions about how to collect data raises a dilemma as to what should be the basis for revision: the test subjects' comments about the difficulties they had encountered as Johnson and Johnson (1975) recommend, heavy reliance on criterion-posttest data as is stressed by Davis et al. (1974), Baker and Schutz (1977), and Markle (1978), or a combination of comments and posttest data as most design models stipulate? This study has been designed to illuminate these and a

number of other issues that should enable instructional designers to better choose from amongst the plethora of recommendations — many of them being offered with little or no research support. Nathenson and Henderson (1977) point to this need for applied research to clarify the procedures:

While developmental testing is obviously an important tool in instructional design, it is nevertheless a tool which lacks a sound empirical base. Thiagarajan (1976), after reviewing American practice concludes: 'We know next to nothing about the variables which contribute to the efficient collection of useful feedback and their translation into appropriate revisions.'

These variables can be described in terms of five interrelated problems, or groups of problems. . .

1. The selection and motivation of subjects for developmental testing.
2. The relationship between performance data and other types of feedback.
3. The development of data-collection systems.
4. The utilization of feedback data in making revisions.
5. The implementation of revisions. . . (p. 9)

Sanders and Cunningham (1974) have made similar conclusions after reviewing the literature related to conducting one-to-one developmental testing sessions:

The present state of the art is crude, consisting of a number of unsubstantiated 'tips' as to how to carry off the procedures. And any inspection of the literature relevant to this topic quickly reveals the inconsistency and lack of agreement among those 'tips'. Some recommend that high ability students be used, others recommend low ability. Some sources argue that students can only clean up semantic and syntactic errors while others insist that the student can make more substantive suggestions concerning sequence, intended prerequisites, etc. Recommendations also vary with respect to preferred level

of student incentive, author behavior in the tryout situation, number of cycles of tryout and revision, and on and on. At present few standard procedures can be recommended with confidence. (p. 31)

A research program was designed to provide a basis upon which standard procedures for obtaining and utilizing student feedback could be recommended based on research. Figure 2 depicts the research activities that were undertaken.

The first component of the research program reviewed the literature of instructional design and the related literature on formative evaluation which offered guidelines for conducting developmental testing sessions. As a result of a content analysis of all of the guidelines, a categorical system was developed and used to classify all of these guidelines. Based on this categorization, independent variables associated with developmental testing were identified.

A questionnaire was then constructed, based on all of the guidelines offered by the literature. Subsequently, a large number of expert instructional design practitioners were surveyed in order to determine the extent to which they agreed with and utilized the various guidelines being endorsed within the literature. Based on the data analysis of the returned questionnaires, a profile of expert recommendations for conducting developmental testing emerged based on the most frequently endorsed procedures.

Using the results of the questionnaire and conflicting recommendations found within the literature, an experiment was conducted to investigate two major variables in developmental testing:

1. the number of test subjects required for testing (i.e., one-to-one versus small group versus both one-to-one and small group testing)
2. the role of the designer and test subject(s) during testing (i.e., active probing and interaction between the designer and test subject(s) geared to identifying problems versus a passive condition not allowing for probing and interaction which could interfere with the validity of the posttest results)

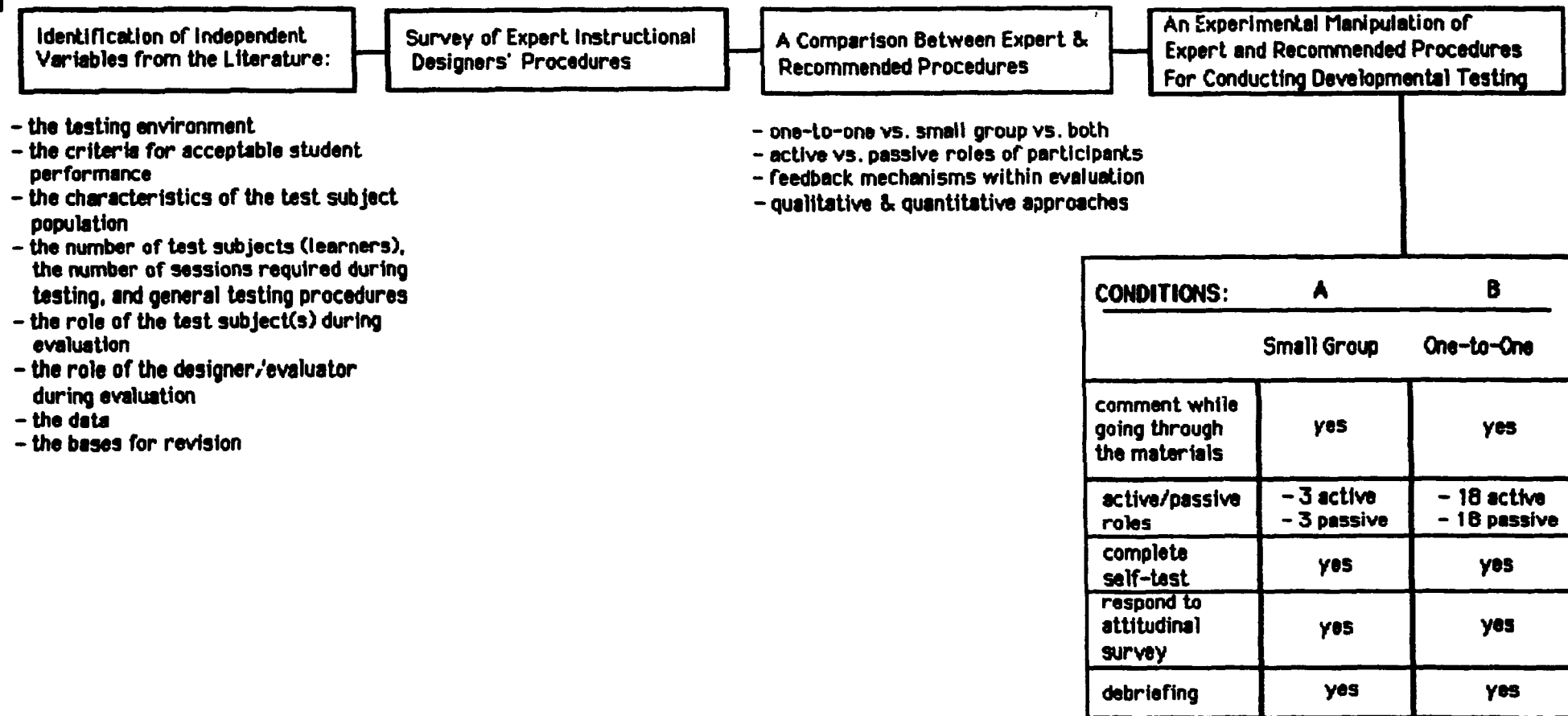


Figure 2. A research programme investigating developmental testing procedures (formative evaluation).

By systematically varying these variables, the effects of the different strategies and roles on the amount and kind of data provided by the test subjects (students) were observed and measured under experimental conditions. A third variable was added to the investigation when it was found that the test subjects utilized in the experiment entered with different amounts of prior knowledge of the material to be learned. As a result, the test subject characteristic of prior knowledge was factored into the analysis. In addition to these variables, a variety of recommended testing procedures were utilized, (i.e., obtaining: test subject (student) comments and suggestions generated while going through the materials, criterion posttest results, attitudinal questionnaire results, and debriefing outcomes). The extent to which each of these testing procedures provided unique feedback data was determined.

A major result of conducting developmental testing is the collection of a large number of qualitative responses that student test subjects make in the form of critical comments and suggestions for improving the materials. A content analysis of these qualitative data was performed in order to categorize the kinds of problems that student test subjects tend to have with self-instructional materials. A similar content analysis was conducted in order to categorize the kinds of revisions or revision prescriptions that student test subjects tended to generate with self-instructional materials. Data were collected in order to determine the most frequent kinds of problems which subjects encountered. Similar data were collected in order to determine the most frequent kinds of revisions/revision prescriptions. Using tabulated frequencies of categorized student test subject problem explications and revisions, the extent to which the following factors quantitatively affect the the kind of data obtained from testing sessions: the role of participants, the number of participants, and the test subject characteristic of either having or not having indicated prior knowledge of the instructional material being testing.

Research Questions

The following are the specific research questions being addressed:

1. What procedures do instructional designers utilize for the design and conduct of developmental

testing?

2. Does varying the assigned roles and the numbers of participants in testing sessions affect the amount of data obtained from the different sources of feedback? and Does varying the assigned roles and the numbers of participants in testing sessions affect the amount of data obtained from the combined sources of feedback?

Sources: a) comments made during instruction (including any comments made while responding to individual self-test items)
 b) debriefed comments
 c) debriefed self-test errors
 d) debriefed attitudinal survey results

3. Does the test subject characteristic of prior knowledge of the material affect the overall amount of comments being made during instruction?
4. To what extent do data from the various sources of feedback provide redundant or unique information? This is a general question addressed by more specific questions:

Specifically: a) After instruction, as a debriefing strategy, to what extent does the procedure of having test subjects go over their comments about the materials again (in retrospect) produce unique or redundant feedback data?

b) To what extent does performance on individual self-test items reflect comments (perceived problems) made during instruction?

c) What are test subjects' perceptions of self-test error causation?

d) After instruction, as a debriefing strategy, to what extent does the procedure of having test subjects review their self-test errors produce unique or redundant feedback data?

e) After instruction, as a debriefing strategy, to what extent does the procedure of having test subjects review their unfavorable attitudinal survey responses produce unique or redundant feedback data?

5. What brings a comment made during instruction to the debriefing agenda, e.g., do subjects just locate and not explain the problem or can they explain and/or remediate the problem?
6. What general and specific kinds of problem explications and revisions/revision prescriptions result from developmental testing and debriefing?
7. Does varying the assigned roles and the numbers of participants in testing sessions affect the amounts of general kinds of problem explications and revisions/revision prescriptions made during instruction?
8. Does varying the assigned roles and the numbers of participants in testing sessions in conjunction with the test subject characteristic of prior knowledge of the material affect the amounts of general kinds of problem explications and revisions/revision prescriptions made during instruction?

By systematically examining the effects of a number of presumably important variables related to designing and conducting developmental testing sessions on the amounts and kinds of data collected, a better understanding of the process of collecting feedback data should emerge from the research.

CHAPTER II

REVIEW OF THE LITERATURE

In this review of the research relevant to the purposes of the present study, two major sources of documentation were utilized: the textbook literature of instructional design and the empirical research literature. First, the textbook literature will be discussed followed by a critical analysis of the empirical research related to the developmental testing of instructional materials.

Instructional Systems Development Models

From a traditional point of view, Dick and Carey (1985) note that instruction has been commonly conceived of as an interactive process between learners, instructors, and textbooks. Knowledge to be acquired by the target population could be found in the textbook. The role of the instructor was seen as a conveyor of that knowledge to the students. A more contemporary conception of instruction has been a systems approach, in which instruction is seen as a process utilizing a number of different interacting components (procedures), which when applied to an instructional goal increases the chance that the goal will be attained. The Dick and Carey design model is fully described in the second edition of their instructional design textbook. This textbook, not unlike others existing in the literature, offers step-by-step procedures for the design, development, implementation, and evaluation of instructional systems.

The guidelines for instructional design as delineated by systems approach models tend to be found in instructional design texts. These texts provide a convenient reference, much like a handbook would, for students of instructional design, practitioners, and researchers. As Dick and Carey (1985) note: "The systems approach models are an outgrowth of over twenty years of research into the learning process. Each component of the model is based upon theory and in most instances research that demonstrates the effectiveness of the component". (p. 4) As such, the textbook literature is a convenient source, as a

departure point, for conducting research on instructional design methodology. For this reason, the first phase of the research was to review major instructional design textbooks (and relevant theses, dissertations, and technical reports) and to select from this sample those texts which offered well-defined procedures for the design and conduct of developmental testing sessions. Appendix A lists the texts and dissertations which were initially reviewed. Appendix B lists the selected texts and dissertations that generated the content analytical data base for the programme of research. The initial phase of the research programme was to conduct a content analysis of all of the procedures so that developmental testing variables could be identified and classified for descriptive research purposes.

Isolation of Developmental Testing Variables

The focus of this overview will be to discuss the different procedures that design-oriented texts tend to use based on their explications of how to obtain feedback data from targetted learners during the prototype testing phase of developing instruction.

Most of the texts reviewed provided a full description of the complete design process. As such, there would be a chapter or section devoted to explaining the rationales and procedures for conducting developmental testing and revision as integral components of the design process. A few of the authors reviewed contributed more extensively by focussing exclusively on developmental testing and revision. Horn (1966) has produced a self-instructional training package for novices which comprehensively explains how to design and conduct testing sessions. Gropper (1975) presents a detailed theoretical analysis of and recommendations for diagnosis and revision. Abedor's (1971) dissertation reviews existing evaluation models and offers a validated model for developmental testing and revision. Lawson (1974) provides a collection of "job aids", based on the work of Sanders and Cunningham (1973), for the formative product evaluator to use for data collection purposes. Nathenson and Henderson (1979) offer a thorough and critical review of the literature. In addition, based on their experience at the Open University in England (a university specialized in distance education), a strategy for developmental testing is presented. Case studies on the effectiveness of the strategy are described. Their own research

on formative evaluation within a distance education course development framework is discussed. Emphasis is placed on positioning, structuring, and utilizing in-text student feedback questions embedded into the instruction.

More common than the above authors' work are the majority of authors who describe a complete instructional systems development model, of which developmental testing is a component. Through the work of authors such as Baker and Schutz (1971); Bjerstedt (1972); Deterline and Lenn (1972); Friesen (1973); Davis et al. (1974); Thiagarajan et al. (1974); Johnson and Johnson (1975); Bell and Abedor (1977); Merrill and Tennyson (1977); Dick and Carey (1978); and Markle (1978) such systems have been developed. The sections of these design texts and the other references listed in Appendix B were examined for the purposes of extracting all of the guidelines and recommendations for conducting developmental testing.

According to Holsti's (1968; 1969) explanations of content analysis methodology, these guidelines and recommendations would be considered as the unit of recording for the content analysis. Each of these verbatim statements was recorded in order to establish a pool of guidelines and recommendations for content analytic purposes. The pool of statements was closely examined in order to group the individual statements into common categories. All of the guidelines were found to fit eight categories representing important variables that authors state as warranting consideration when designing and conducting developmental testing. All of the authors represented in Appendix B recommended guidelines which focused, in differing ways, on these variables:

1. The criteria for acceptable student performance.
2. The number of test subjects (learners), the number of sessions required during developmental testing, and general testing procedures.
3. The characteristics of the test subject population.
4. The role of the test subject(s) during developmental testing and revision.
5. The role of the instructional designer/evaluator during developmental testing and revision.
6. The data.
7. The basis for revision.
8. The testing environment.

The review of the textbook literature to follow has been organized to reflect the recommendations by authors of instructional design texts with regards to these eight identified categories or variables.

Developmental Testing Variables Discussed in Textbooks

The Criteria for Acceptable Student Performance

A number of authors have addressed the issue of defining standards for instructional effectiveness. The answer to the question: when is the instruction good enough?, implies the point at which the designer/evaluator terminates the iterative cycle of developmental testing and revision. Generally, the authors approached defining criterion performance in terms of test subjects' achievement on the posttest. For example, a score of 70 - 80% on the posttest can be taken as evidence that the preceding instruction was effective. Abedor (1971) suggests using a mastery learning criterion, 80% or more of the student tryout group should achieve 80% or higher on the posttest. Johnson and Johnson (1975) recommend that a first draft should elicit 70% correct responses from 70% of the tryout learners. Baker and Schutz (1971) state that evaluation should continue until variability in the nature of the instructional treatment is eliminated (in reference to sufficient control of the students' responses so that different students are not inferring different tasks from the instructional material). Deterline and Lenn (1972) indicate that the test subject must be tutored, if necessary, until the specified objectives have been successfully attained. Thiagarajan et al. (1974) recommend developmental testing until a pattern of consistent results are being obtained.

Some authors suggest a criterion based on errors being made on embedded practice items within the instructional material itself. This was referred to as "frame error analysis" during the programmed instruction movement of the 1960's. Programmed instructional materials presented "frames" to learners, e.g., a diagram or a few sentences describing a single concept. Immediately after each frame was presented a discrete, overt response was required to be made by the learner. Friesen (1973) advocates a 7% frame error rate criterion, (i.e., if 7% or more of the students have trouble on an embedded item (a frame) then the practice item should be revised). Similarly, if 7% or less of the students make use of an item then its validity should be questioned or the item eliminated. Dick and

Carey (1978) also advise inspecting embedded practice item data although they do not provide a specific criterion. Bell and Abedor (1977) apply frame error analysis methodology to the individual posttest item. If 20 - 30% or more of a group of between four and six test subjects respond to a particular test item incorrectly then that item should be discussed by the developer with the group after developmental testing in a debriefing context.

Some authors, in addition to posttest performance data, suggest the additional criterion of utilizing affective responses to Likert-type scaled items and open-ended items from an attitudinal survey. Abedor (1971) has produced a validated attitudinal survey to which he suggests that no more than 20% of the groups' mean responses on the post instruction attitudinal survey should indicate dissatisfaction. Similarly, Bell and Abedor (1977) advocate that if 30% or more of the group of test subjects deviate two or more points away from the ideal on a Likert-type scaled item then the material should be revised. Merrill and Tennyson (1977) also recommend using student ratings but in a different manner from the traditional survey. Beside each item on the practice and test displays, students are asked to rate the item as to whether the respondent felt sure, unsure, or guessed on the item. In addition to the confidence rating, these authors suggest comparing pretest/posttest performance at both the individual item level of error analysis and the total test score level.

Most authors recommend attending to error patterns made on the embedded practice items and/or posttest as possible indicators of faulty instruction. Gropper (1975) is a strong proponent of the use of posttest performance data as the cornerstone of the student test subject data base. This posttest, to be a valid and reliable indicator of instructional effectiveness, must be technically well-designed and comprehensively sample the attainment of the instructional objectives. When test items (often criterion-referenced) are specifically aligned with those particular segments of the learning material which offer instruction for the knowledge and skill domains on which learners are being tested, appropriate diagnostic activities can be conducted. Baker and Schutz (1971); Bjerstedt (1972); Deterline and Lenn (1972); Dick and Carey (1978); and Markle (1978) recommend the keying of test items directly to the instructional content in order to pinpoint the location of the test subject's problem.

Theoretically, another set of criteria for terminating developmental testing could be based upon the behavior of the test subjects while going through the instructional materials. One criterion might be the frequency and kind of comments (indications by the students that they perceive problems within the materials) made while trying to learn from the instruction. After repeated testing incorporating revisions based on these comments, criticisms should decline and comments should become redundant. None of the reviewed authors specifically suggested this kind of criterion.

The Number of Test Subjects (Learners), the Number of Sessions Required During Testing, and General Testing Procedures

The size of the suggested test subject population depends to a large extent on how any given author conceptualizes the developmental testing session(s). If prototype testing is seen as an opportunity to use intensive, clinical, probing activities that tend to generate qualitative feedback data, then an author will likely recommend one-to-one sessions (the developer/evaluator sitting with the test subject as he/she goes through the material). On the other hand, small groups (e.g., six test subjects) will likely be suggested by authors who emphasize data gathering during the instructional session with emphasis on posttest and attitudinal survey data (quantitative data). Horn (1966); Bjerstedt (1972); Deterline and Lenn (1972); Johnson and Johnson (1975) recommend conducting clinical one-to-one sessions only. Abedor (1971); and Bell and Abedor (1977) recommend using solely small group sessions. Dick and Carey (1978); Friesen (1973); Thiagarajan et al. (1974); Lawson (1974); and Merrill and Tennyson (1977) recommend both one-to-one and small group.

Authors vary as to the recommended number of one-to-one sessions to be conducted. Horn (1966) suggests that the number of students will vary from one to 12 with usually enough information obtained from three students; Bjerstedt (1972) suggests one or more students while emphasizing the number of students being less important than the detailed analysis of how well the communication penetrates; Deterline and Lenn (1972) advise testing with three students; and Johnson and Johnson (1975) advocate two or three students. Of those authors endorsing both one-to-one and small group testing Dick and Carey (1978) recommend one-to-one with two or more students; Friesen (1973) advises

one-to-one testing until "a success factor" is reached; Thiagarajan et al. (1974) suggest four or five one-to-ones; and Merrill and Tennyson (1977) advise one-to-one testing with one subject however a second subject should be used if numerous changes result from the first.

Authors vary as to the number of small groups to be employed. Abedor (1971); Dick and Carey (1978); Friesen (1973); Bell and Abedor (1977); and Merrill and Tennyson (1977) all suggest that one group session might be sufficient while Thiagarajan et al. (1974) suggest as many as four or five small group sessions might be conducted. The number of test subjects within a small group is also unclear based on the recommended guidelines in the literature. Abedor (1971) suggests nine to 12; Dick and Carey (1978) 10 to 20 in a group; Friesen (1973) suggests six; Bell and Abedor (1977) recommend four to six students; and Merrill and Tennyson (1977) advise using 15 in a group. Baker and Schutz (1971), suggest, as a general guideline, to avoid using an extremely small or extremely large number of learners. They indicate that there is no "exactly right" number. However, they state that two or three test subjects would be insufficient because there would be less chance that reliable information about the usefulness of the product could be obtained.

There is extreme variability among the guidelines in the literature with regard to the number of subjects and kinds of sessions. The question of how many test subjects should be used can be related to the profile of the test sample. An author who suggests one-to-one sessions is also likely to suggest using more than one session with different learners until learning criteria (as discussed in the previous section) and an appropriate sample of the target population has been covered. Presumably, a major reason for using more than one test subject follows from an assumption that there will be some variability in the data across subjects which should be taken into account for revision purposes. Another reason may be to enable the designer to validate, on another test subject(s), changes that have been made as a result of data gathered in the previous session.

There is much less variability among the recommendations in the literature with regard to debriefing test subject(s) after the testing session. The majority of authors endorse discussing the data obtained through the conduct of developmental testing with the test subject(s). This general procedure

presumably finalizes the results for a given testing session by providing test subject(s) with the opportunity to add, elaborate, or modify the obtained feedback data.

The Characteristics of the Test Subject Population

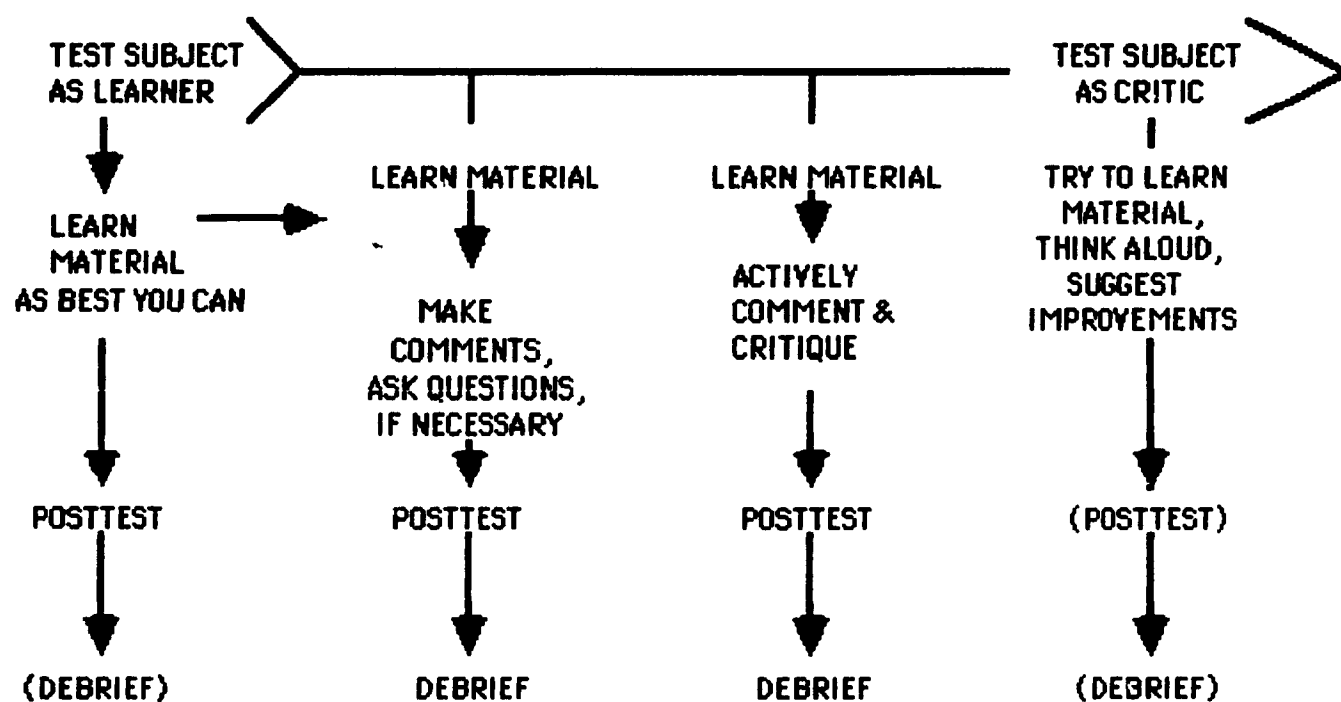
There is general agreement among authors that the test subject population should be representative of the target population for whom the prototype instruction was designed. Deterline and Lenn (1972); Friesen (1973); and Davis et al. (1974) specifically recommend using student test subjects who have reached the point in the course for which the lesson being tested would be appropriate. Bjerstedt (1972) qualifies this general subject characteristic by recommending that the students should not be too timid or docile. Thiagarajan et al. (1974) place less importance on test subjects being representative of the target population. They suggest that test subjects must be enthusiastic and highly verbal as these characteristics will enhance subjects' ability to verbalize problems. Johnson and Johnson (1975) have proposed that anyone can be utilized as a test subject. Presumably, they would expect the designer/evaluator to exercise common sense in the selection of test subjects. Some authors suggest stratifying the test subject sample of learners according to "high, average, and low" levels of ability or achievement. Horn (1966) and Dick and Carey (1978) recommend this stratification procedure in order to insure this heterogeneous mix. Most authors will encourage heterogeneity within the test subject population although they do not emphasize systematic stratification. Abedor (1971) advises using subjects of varying abilities as differentiated by S.A.T. scores and Bell and Abedor (1977) recommend using subjects of varying ability. Markle (1978) suggests that through developmental testing specific pre-requisites will emerge regarding the type of subjects to use although a general description will provide a good idea of general pre-requisites such as reading level or I.Q. Lawson (1974) has recommended using poorer learners. He has assumed that if "low" ability/achieving students can successfully complete the materials so will the other students. Henderson and Nathenson (1977) recommend, based on their experience in developing distance education courses at the Open University, that if the target population is extremely heterogeneous in

nature, test subjects sampled to represent the target population must be stratified by way of their entering behavior on a pretest of the learning materials.

From this review it is clear that authors have not specifically addressed the issue of how similar the test subject sample should be to the target population. Authors appear divided on the issue of along which dimensions should the two populations coincide. As long as these remain unresolved issues it will be unclear as to how to select student test subjects. As the review reveals, systematic stratified sampling, pretest data, pre-instructional data such as S.A.T. scores, I.Q. scores, and reading comprehension level scores, or personal learner characteristics of being articulate and extroverted all have been recommended by different authors.

The Role of the Test Subject(s) During Developmental Testing and Revision

Authors' descriptions of the roles that participants (student test subjects and the developer/evaluator) should play during developmental testing and revision precisely illustrate how the authors generally conceive the developmental testing process. Figure 3 presents a variety of roles which have been proposed within the literature of instructional design for student test subjects. These roles can best be visualized along a continuum. At one extreme the test subject is seen as a "traditional student" — acquiring the appropriate knowledge and skills to the degree indicated by posttest results. Markle (1978) and Baker and Schutz (1971) subscribe to the traditional student role for a test subject(s). In this case, where the test subject assumes a passive role, the primary data upon which the developer/evaluator bases decisions to continue to developmentally test and make revisions are the posttest results. At the other end of the continuum is the active test subject who is encouraged to think aloud, make critical comments and suggest revisions while working through the instructional materials. Thiagarajan et al. (1974) and Bjerstedt (1972) might be placed at this end of the continuum. In between these two extremes are test subject roles which range from being permitted to make occasional comments to a more active role in which the test subject is encouraged to make critical comments. Friesen (1973) recommends that students be allowed to determine the causes of their



() = OPTIONAL

Figure 3. Role of test subject

problems when the developer notices a problem. Horn (1966) and Merrill and Tennyson (1977) suggest that student test subjects should be allowed to write or express any comment they want to make with regards to the problems they are encountering and, in addition, to rate the level of confidence they have on their responses to individual posttest items. Abedor (1971); Beil and Abedor (1977); Deterline and Lenn (1972) suggest that a post-instructional debriefing session will facilitate the active role that the test subject or group of subjects is encouraged to play. In some cases, authors such as Dick and Carey (1978) and Johnson and Johnson (1975) advise that test subjects be requested to write their comments either on a comment sheet or on the instructional materials; or these comments are being recorded by the developer/evaluator either on a master copy of the instructional material or by using a tape recorder.

Reflecting these different roles that test subjects might be asked to assume, test subjects may be instructed (at the onset of a developmental testing session) to, e.g., "Go through the instruction as best you can"; "Ask questions only when you are really stuck"; or "Comment on points of difficulty as you work through the materials". In general, suggested student test subject roles range from that of test-subject-as learner to test-subject-as-critic.

An important point already raised is the relationship between test subject activity during testing and the emphasis placed on the results of the posttest. Obviously the more that the student interacts with the developmental tester (resulting in more explanation and on-the-spot revisions, etc) the less are the posttest results a reflection of the instructional materials. Consequently, it is recommended, by those who advise this kind of active intervention, that precise descriptions of changes made during testing be noted.

The Role of the Instructional Developer/Evaluator During Developmental Testing and Revision

Similar to that of the student test subject role, the developer/evaluator role is described quite differently by various authors. The roles of the participants tend to complement each other; if an active role is prescribed for test subjects then an active role will be prescribed for the developer/evaluator. Figure 4 illustrates a continuum of proposed roles. Recommendations range from a tester assuming the role of a session supervisor (e.g., providing instructions to the test subjects for task orientation purposes, recording learning times, and analyzing test results), as is suggested by Baker and Schutz (1971); Davis, et al. (1974); and Markle (1978) through the passive role of an observer (e.g., recording the reactions of the test subjects) as is suggested by Thiagarajan et al. (1974), to a very active role in which the tester becomes an active intervener either as a tutor (when the test subject encounters difficulty) as Dick and Carey (1978) and Bjerstedt (1972) recommend or as someone who probes the test subject for comments and suggestions as is suggested by Horn (1966); Deterline and Lenn (1972); Lawson (1974); and Merrill and Tennyson (1977). Johnson and Johnson (1975) advise discussing potential revisions, between the participants, in order to revise on-the-spot. Authors' guidelines within the literature reflect a great deal of variability as to the degree of intervention to be initiated by the tester. Abedor (1971); Bell and Abedor (1977); and Friesen (1973) propose a rather reactive role — respond succinctly to test subject questions; others suggest that the tester should be regularly intervening through actively probing the test subject with regards to locating, defining, and resolving possible problems perceived within the instruction.

There is general agreement by authors that, immediately after the test session (following a short break), the test subject(s) be asked to reconvene for a debriefing session. This post-instructional interview situation is seen as a finalizing data collection procedure for the feedback being generated. Comments during developmental testing are being elicited "on-line" as the test subject(s) are trying to both learn and criticize simultaneously. What may have been perceived as a problem during developmental testing may no longer be considered to be problematic afterwards. Conversely, new problems may be uncovered during debriefing that were not alluded to during developmental testing.

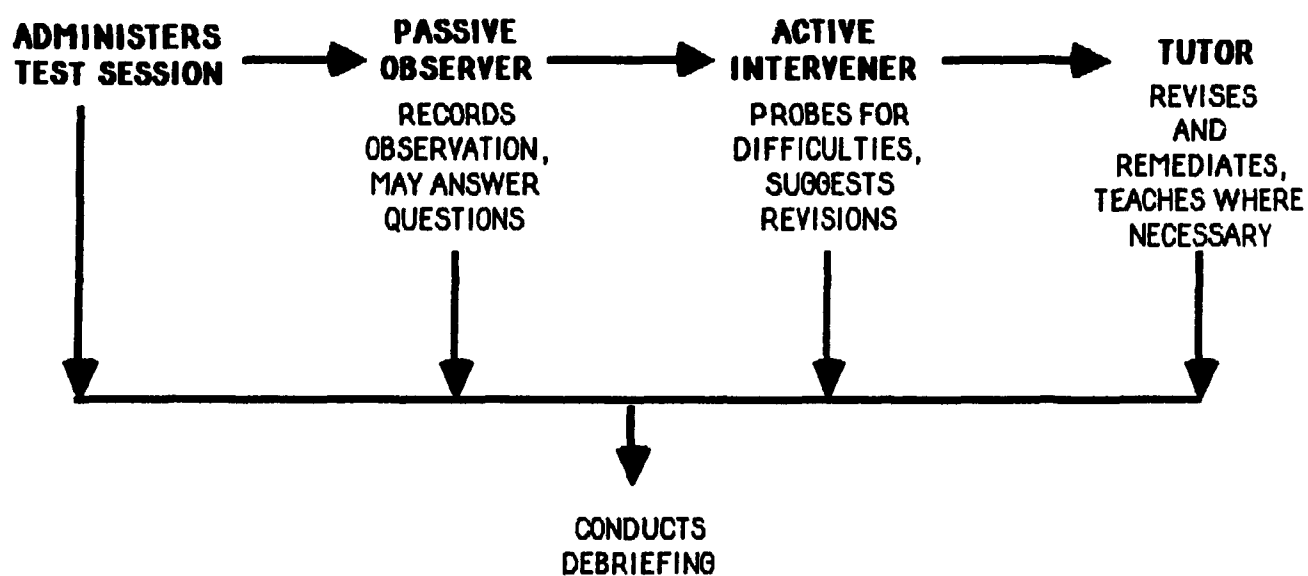


Figure 4. Role of developer/evaluator

Gropper (1975) offers specific guidelines for conducting debriefing sessions. Thiagarajan (1986) provides a comprehensive synthesis of debriefing activities and their underlying rationales. Generally, it is recommended that the developer/evaluator should set the agenda, probe for difficulties subjects encountered, and summarize comments.

It should be noted that the term developer/evaluator is being used as if one person were carrying out all of the activities. Some authors have recommended a differentiation of roles in order to control for potential bias. These authors suggest that the designer should not be the evaluator. An external evaluator supposedly will be more open and receptive to accepting critical feedback and suggesting revisions to the instructional design team.

The Data

Authors reviewed universally endorse the employment of a posttest, although, as previously indicated, emphases on how the posttest data will be utilized varies. Most authors also prescribe the administration of an attitudinal survey in order to tap affective reactions to the instruction.

Those authors who advocate active, clinical one-to-one sessions have emphasized gathering student test subject comments and suggestions. Various methods have been offered in order to record this kind of qualitative data: tape recording the session, subjects either writing their comments on the instructional materials or on a revision comment sheet or having the developer perform this task, the developer/evaluator noting observations made during the test session, writing proposed revisions on a master copy of the instructional material, notes recorded on the instructional assistance required by test subjects as they were going through the material (i.e., the tutoring), notes on the non-verbal behavior of the test subjects, student comments made during debriefing, etc. Results obtained from the embedded practice exercises ("frame" error data), attitudinal survey data, and the pre and/or posttest data all can be considered to be quantified data.

Authors recommending small group testing (either as the method for testing materials or in combination with one-to-one testing) often will place less emphasis on the developer/evaluator

interactions with the test subjects during testing although attending to problems is permitted. These authors acknowledge the difficulties in using a clinical orientation with groups of test subjects as they are trying to learn from the materials — simultaneous tutoring of different subjects at different locations within the material (proceeding at different paces) is not an easy task. Test subjects in group-based developmental testing are often asked to write down their problems, comments, and suggestions in order for them to be discussed afterwards. This is presumably why debriefing is endorsed as a procedure within small group testing. During the debriefing, difficulties that test subjects have located, constructive criticisms, and suggestions all can be discussed during group debriefing.

The Basis for Revision.

Revision can be conceptualized as a diagnostic and prescriptive process, based on developmental testing, in order to create instructional materials that meet set performance criteria. The reason, in practice, for conducting developmental testing sessions is to obtain data for the purposes of revising the instruction; therefore procedures for developmental testing and revision interact. For example, it has been suggested that student test subjects may be asked to actually express (verbally or in writing) revisions during the developmental testing and/or debriefing session. In a situation like this, it becomes artificial to dichotomize developmental testing and revision. "Revision, after all, is merely the application of development techniques at a different point in time" (Groppe, 1975, p 187). Dick and Carey (1985) recognize this dilemma in the second edition of their text by informing readers of their rationale to present two separate chapters: one on testing and the other on revision. They state that:

A rather arbitrary division of content has been made between this chapter and the next. We typically think about formative evaluation and revision of instructional materials as one major step. However, for the sake of clarity and to emphasize the

importance of re-examining the whole instructional design process when instructional materials are to be revised, we have separated the design and conduct of the formative evaluation study from the process of revising the instructional materials (p. 198).

For similar reasons, the focus of the research being reported on in this dissertation is on designing and conducting developmental testing sessions. However, when specific prescriptive activities for development testing interact with revision activities to the point where they become indistinguishable then revision is addressed. In this context, the basis for revision will be reviewed and addressed throughout the research being reported as opposed to the process of actually revising.

How an author views developmental testing obviously affects how that author will conceive the bases upon which the instruction is to be revised. Different kinds and amounts of information to be obtained from a variety of proposed sources of feedback have been suggested in the literature. Within the broad context of revising based on the feedback obtained from implementing developmental testing procedures, Figure 5 outlines proposed procedures which, in effect, describe the bases of revision. These procedural strategies emerged from the same textbook literature as the guidelines for developmental testing. Authors' recommended strategies for revision are as variable as guidelines for developmental testing, and generally take the form of revision algorithms, heuristics, rules, and guidelines.

The first component of the bases of revision model reflects how problems might be identified. All of the reviewed authors recommend some way to "flag" errors or problems within the instruction. The examples presented below the boxes reflect a range of suggestions within the sampled literature. "Frame errors", student nonverbal behavior, and posttest results all have been proposed.

Next, the reviser must determine how important the identified error is. A useful guideline is to apply the standard of what will be expected on the part of the learner in terms of criterion performance. Tolerance for errors on the part of the reviser (i.e., the threshold at which an error will be attended to) can be related to the previous discussion on the criteria for acceptable student performance.

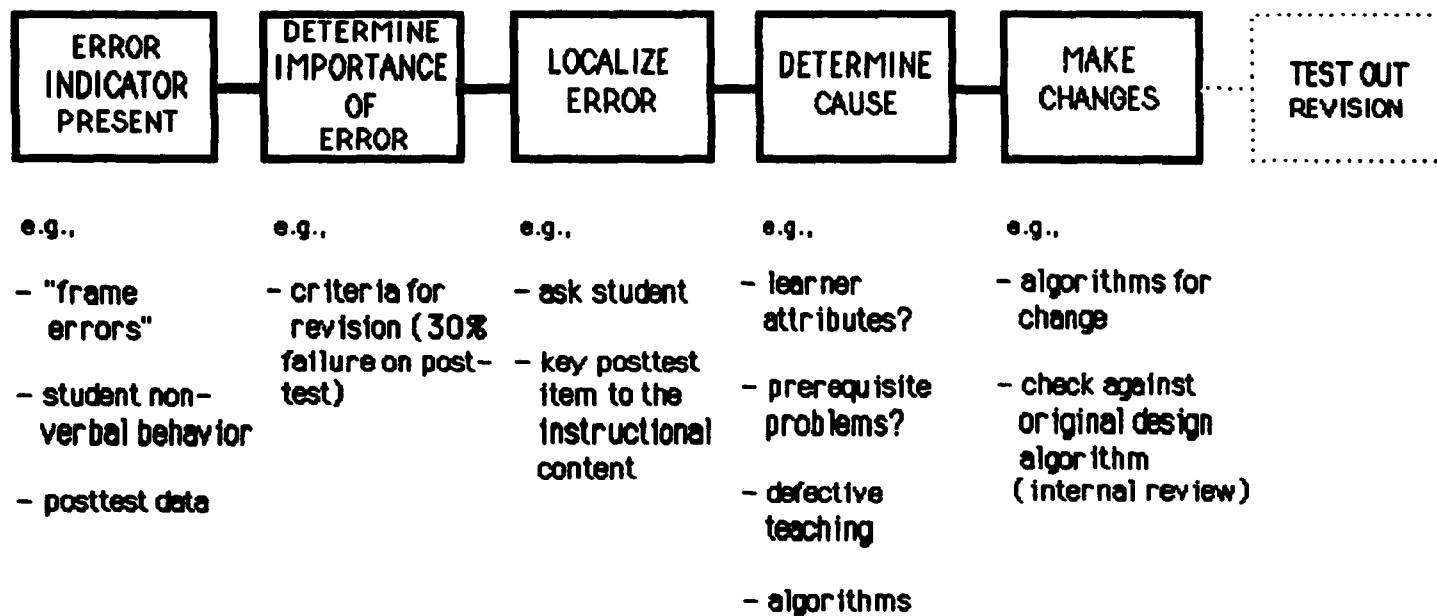


Figure 5. The bases of the revision process.

After an error has been judged as warranting further attention, the next step would be to locate the error geographically within the instruction. Authors such as Horn (1966); Abedor (1971); Deterline and Lenn (1972); Friesen (1973); Johnson and Johnson (1975); Bell and Abedor (1977); and Dick and Carey (1978) rely upon the student test subject(s) to locate the problems encountered when using the materials. Baker and Schutz (1971); Bjerstedt (1972); Deterline and Lenn (1972); Dick and Carey (1978); and Markle (1978) advise locating the problem via the posttest items which have been directly keyed to the instructional content. Bjerstedt (1972) and Dick and Carey (1978) couple the keyed test approach with test subject problem localization as their methodology. Student test subjects are instructed to pinpoint problems/errors by either verbally indicating problems to the developer/evaluator, writing the encountered problems down on a revision sheet designed specifically for this task, or by writing on the instructional materials themselves.

After problems have been explicated, the reviser must hypothesize the cause(s) of these identified problems or errors. Markle's (1978) position is that developmental testing is the prime vehicle through which learner pre-requisites and general attributes such as reading comprehension level and/or I.Q. will empirically emerge. Brethower (1965); Horn (1966); Bjerstedt (1972); Friesen (1973); Thiagarajan et al. (1974); Johnson and Johnson (1975); and Dick and Carey (1978) suggest that the cause of instructional failure might be defective teaching. When such defects are encountered during developmental testing, the instructional designer has the option of overriding the defective instruction by on-the-spot generated teaching or tutoring. Ideally the developer/evaluator, in this case, would be able to supplement ineffective instruction with effective instruction ("online revision") and, when unsuccessful at this attempt, continue to do so on a "trial and error" basis until the test subject(s) comprehends or acknowledges that the defect has been remedied. As this kind of revision process is dynamic, it is frequently suggested that the entire session be tape recorded. This allows for a transcription of the "tutoring" to be used as a basis for revision. The causality of problem/error is inferred through successful remediation. Actual changes that are made this way often occur when authors prescribe clinical one-to-one sessions. Revision can also occur after developmental

testing (i.e., after the posttest and attitudinal survey have been completed) during the debriefing session. This procedure is frequently recommended by those advocating small group testing.

There are a few authors who deviate from the revision methodologies discussed. Basically, those discussed up to now have opted for a recycling process basis for revision — recycle back through the original instructional design procedure in order to insure that the resulting product is congruent with specified design procedures. Student test subject data should facilitate identifying inadequacies and incongruencies. Revisions are then inferred from the data gathered from developmental testing. Dick (1968); Deterline and Lenn (1972); Robinson (1972); Gropper (1975); and Merrill and Tennyson (1977) are atypical of most authors in that they have developed revision algorithms to be used in order to specify problem/error causation and prescribe revision using appropriate developmental testing data. Presumably the use of a valid revision algorithm reduces the need for a reviser to infer what needs to be changed and how it should be changed. Research by Dupont (1980) has been conducted in establishing the reliability and validity of a revision algorithm for audio-visual learning materials. Such demonstrations across various media should help reduce or eliminate the level of subjectivity which presently characterizes much of the revision process.

It is apparent that revision, much like developmental testing, spans a continuum of roles for the participants of testing sessions. For example, test subjects may be permitted to serve in the traditional role of student, trying to learn the material as best they can by attending and working through it until learning is evidenced through a comparison of pre and posttest results. At the other end of the continuum a subject assumes the role of "colleague" with the developer/evaluator. Student test subject "colleagues" are seen as partners in the development phase — somebody who can provide valid and reliable comments/recommendations about instruction that has failed to provide the conditions for learning for particular test subjects. As noted, these "colleagues" may be asked to actually revise instructional materials so that individual differences with regards to learning the subject matter can be attended to and minimized for the target population to whom the instruction will ultimately be delivered.

The Testing Environment

Few authors have commented on the physical environment at which developmental testing and debriefing occur. Only Horn (1964; 1966) seems to have seriously considered this environment and offered recommendations. His suggestions are obvious. Test subjects should be given pens rather than pencils so that erasures of comments and revisions can still be legible. Put the test subjects at ease by telling them that it is the instruction that is being tested and not them, etc. Others (e.g., Paulson, 1969; Abedor 1971) have incorporated Horn's guidelines with regards to the testing environment. Henderson and Nathenson (1977) recommend that a developmental testing session should take place in an environment very similar to the actual environment in which the material will eventually be used.

A Summary of the Textbook Literature of Guidelines

There is a certain amount of agreement, but by no means complete agreement, among the authors reviewed on particular procedures to be utilized for the design and conduct of developmental testing sessions. The use of criterion-referenced posttests, of test subjects who are representative of the target population, and of prototype instructional materials rather than a finished product are examples of where there is considerable agreement.

There are major differences in how authors conceptualize testing methodology. Illustrative are the wide range of recommendations for the number of test subjects to be sampled and through which selection process, the size of the group in which these subjects are to provide feedback (i.e., one-to-one, small group; or under both of these conditions), the kind of feedback data to be obtained and from which sources, and the roles that participants (test subjects and the developer/evaluator) should play.

There is typically a paucity of advice linking the guidelines for developmental testing and revision guidelines. Authors offer little help in suggesting how to translate data gleaned from developmental testing into effective revisions. As Dick (1980) notes, as an instructional design textbook author, a

teacher of instruction design, and a monitor and frequent contributor to the research literature on formative evaluation:

The greatest limitation in formative evaluation today (is) the dilemma of what to do after a problem has been detected in instruction. Nearly all instructional writers have indicated that after the data have been collected and summarized in a formative evaluation, the designer should 'revise appropriately'. However, in most instances, designers have already used their best knowledge of how to design the instruction, and therefore it is not always apparent what 'revise appropriately' would be (p. 5).

Critical Analysis of Empirical Studies

A number of reviews of the literature related to developmental testing have periodically been conducted and reported by a number of journal authors. Tyler (1942); Rose and Van Horn (1956); Scott and Yelon (1969); Baker and Alkin (1973); Sanders and Cunningham (1973); Komoski (1974); Smith and Murray (1975); Thiagarajan (1976); Nathenson and Henderson (1977); Thiagarajan (1978); Dick (1980); Kandaswamy (1980); Cambre (1981); Sarapin (1981); Reigeluth, Van Patten, and Doughty (1981); and Weston (1986) have reported reviews of the literature. Empirical studies on developmental testing variables, and reviews of them, have aided researchers working towards the ultimate goal of validating testing guidelines through a better understanding of the variables contributing to more effective instructional design practice.

The following review of the empirical literature related to developmental testing variables, and literature germane to the topic, utilizes empirical research on formative evaluation pertinent to four specific research traditions or paradigms:

1. Internal review
2. Field testing
3. Learner verification and revision (LVR), and

4. Developmental testing

Each of their contributions related to developmental testing will be presented and discussed.

Internal Review

As previously discussed, there are proponents of internal/expert review procedures who suggest that the use of test subject feedback is unnecessary. In the context of this study, such a recommendation must be taken seriously. It is recognized that conducting formative evaluation is an expensive and time-consuming effort. From a marketing perspective, publishers of instructional materials can cite the competitive nature of the market place as a major reason why it is difficult for them to support product testing — another product from another publishing house with a different author writing similar instructional material might corner the market and reduce sales for a publisher committed to pre-publication testing procedures that would delay publication. Ideally, if the exclusive use of internal/expert review could be demonstrated as being an effective strategy for review and revision purposes, then gathering data from test subjects would not be necessary. For this reason, empirical studies using well-defined models for conducting an internal review will be presented and discussed in relation to learner data.

Montague, Ellis, and Wulfek (1983) offer the following as an operational definition of internal review:

The designer and a subject matter specialist perform an internal review of preliminary instruction for accuracy and completeness. They judge the adequacy of the instructional sequence, the motivating capability, and the support for student learning (p. 11).

This definition is based on a description of the Instructional Quality Inventory (IQI), as a curriculum evaluation and design tool, found in Merrill, Reigeluth, and Faust (1979). The IQI provides specific guidelines for the internal review of instruction under development or already developed. It replaces

the data that would be obtained from using tryout subjects. It is of interest to note that the authors state that: "These procedures have undergone field testing and extensive revision at our center" (p. 11). It does appear that "users" of the guidelines did provide valuable feedback for improving the inventory. The authors do indicate that there are limitations on what can be expected from using the inventory: "The IQI is intended for use with cognitive and psychomotor instruction, and is appropriate for evaluating most forms that those sorts of materials take" (p. 11). From this statement, it seems that the IQI is not equipped to deal with affective/motivational aspects of the instruction. The authors candidly admit that: "A limitation is that the IQI does not address either adapting instruction to students or the sequencing or structuring of content" (p. 13). These are critical instructional design decisions.

Burkholder (1982) utilized the ISDP (Instructional Strategy Diagnostic Profile), developed by Merrill, Richards, Schmidt, and Wood (1977), in an experiment geared to exploring its effectiveness as an aid in conducting an internal review. This profile consists of a set of questions and ratings to facilitate the analysis of the instruction for purposes of generating prescriptions or revisions. The process can be applied to both existing materials or materials yet to be developed. There are four major components aligned with the profile. Using a task-content matrix, learning objectives and their related test items are classified according to the content to be learned and how it is to be learned. Next, both test items and objectives are evaluated as to their degree of similarity/consistency. Then terminal objectives are analyzed in terms of final performance levels required by the objective and compared with the needs and demands of the curriculum. Instructional episodes are finally classified according to their mode. According to the underlying propositions of the ISDP, instruction can take on four different presentation modes. A learner can be instructed by being presented: 1) with generalizations, 2) with examples, 3) with questions about generality (Generality Practice) or 4) with questions associated with examples (Practice). The profile allows an instructional developer to monitor the interactions of the above four major components in order to judge the intended instruction qualitatively.

Utilizing a rigorous experimental design, Burkholder (1982) does provide convincing evidence as to the profile's ability to generate improved materials. Three experimental groups were assigned to

treatments which utilized three variations of a self-instructional learning package designed to teach a basic instructional media course to students at the Pennsylvania State University. The module had had a history of poor student performance. The profile was used to generate two different versions of the original product. Levels within the design employed the original material, a revised version of the original rated (using the profile) as having improved consistency, and a version rated as having improved consistency and adequacy. Marked improvement over the original was found on such dependent variables as: student performance, student confidence in what they had learned, and attitude towards the ISDP revised learning materials. Unfortunately, the author does not describe how the original materials were designed, e.g., using what kind of design model? He does cite substantial research supporting the theoretical basis of the ISDP and notes the empirical approach, utilizing data generated from student test subjects, as being a relatively undefined technique.

Choi, Merrill, Callahan, Hawkins, and Norton (1979) conducted a similar experiment to Burkholder's except that a different format of instructional material was utilized. Two organic chemistry texts from an undergraduate Brigham Young University chemistry course were used as the stimulus material. The ISDP was utilized to rate both texts. This application of the profile resulted in one text receiving a significantly higher rating than the other. Only partial support was obtained for a hypothesis stating that the higher rated text would generate statistically higher student performance ($t = 1.3745, p < .10$). A hypothesis stating that there would be a significant linear relationship between ISDP ratings and student test performance as related to the two text books was not supported by the data. The most interesting finding did indicate that when the Chemistry Department ultimately selected the text with higher ISDP ratings, significantly higher student performance was observed. Posttest scores were generated from the National Standard Organic Chemistry test, the 1974 ACS Cooperation Exam in organic chemistry. From the authors' discussion and recommendations sections, it was perceived that to be truly effective, there must be a one-to-one correspondence between test content and test items. Because a standardized exam was employed, and because the textbooks were not of a self-instructional format, which would have had this built in one-to-one correspondence, the correspondence was not completely met. This could explain the lack of a linear relationship between the

ISDP ratings and student performance. This aspect of the design of the material might also explain why Wood, Richards, and Merrill (1976) have found a high correlation between student performance and ISDP ratings of materials. If the rating is to be truly effective, test items must correspond with content.

Similar to having instructional design experts and content experts formatively evaluate instructional material within an internal review research framework, Rothkopf (1963) was interested in looking at teachers in order to determine if they could function as effective reviewers outside of their content areas. After having completed a three week course in programmed instruction, 12 high school teachers and principals were asked to predict the effectiveness of seven self-instructional arrangements of anthropological information. The rank order correlation between the subjects' predictions and the observed effectiveness (target population posttest scores) was $-.75$. This study suggests that merely "inspecting" materials is not enough to evaluate the effectiveness of materials.

Smith (1978) was also interested in investigating teachers' ability to predict student motivational and achievement variables. Her study compared teachers' evaluations of an educational television script with subsequent student audience perceptions and achievements. Four teachers rated the scripts of a television series intended to reduce sex-role stereotyping in occupational related activities for nine to 12 year olds. The student sample consisted of 666 fourth, fifth and sixth graders from seven school districts in the Los Angeles County area.

A comparison of teacher and student evaluations was conducted along a number of dimensions but the most interesting were teacher ratings of the scripts in terms of their interest level, language level, and their ability to convey factual information. Students were asked to provide data, after viewing the films, on their interest levels, comprehension of the factual information, and their comprehension of what the objectives of the films were.

... results indicate a moderate and practically interpretable relationship between teachers' ratings of student liking, interest, and 'getting the point' with students'

'likability' ratings of production segments and answers to questions about 'facts' in the segments. Similarly, teachers' ratings of 'language' level seem to predict the same responses, but mainly in a direction of an inverse relationship between simplicity of language and high factual comprehension. Clearly, however, the teacher ratings had no utility in predicting responses to questions about the instructional objectives (p. 51).

Expounding on Lumsdaine's (1965) differentiation of internal and external sources and characteristics of formative evaluation, Cunningham (1971) describes the internal characteristics of materials, such as the content to be covered, readability, and emphases of the particular topics covered in the instruction. This is contrasted with external characteristics of materials which reflect the specific, observed effects of the materials on the target population and various 'significant others' such as parents, teachers, etc.

This distinction is similar to that made by Scriven (1967) between pay-off and extrinsic evaluations . . . An intrinsic evaluation would be concerned with the internal characteristics of the instructional materials while a pay-off evaluation would concentrate on the external effects of the materials (p. 5).

Cunningham (1971) proposes a classification system for information obtained from formative evaluation. The first type, internal information, is collected to describe the materials and to obtain critical appraisal from a number of possible sources. Secondly, external information is obtained as to the effects of the materials on student, teacher, and any other relevant "others". The third category deals with contextual information — information about the characteristics of the teacher, the student, and the setting in which the instruction is to be used. This type of information delimits the range of the target population for the materials.

From the above discussion, it could be argued that internal/expert review procedures address what has been termed as internal information or characteristics and intrinsic evaluation. In this context, internal/expert review procedures and student test subject empirical data (external review) are seen as being complementary to one another. The expressed limitations, on the part of instructional researchers who have developed internal review procedures, and the limiting conditions under which an evaluator uses the strategy must be taken into consideration.

Internal review strategies tend to be analogous to a checklisting tool that requires the developers to recycle back through all of the components of the design process in order to look for inadequacies and misalignments of design decisions and implementations. It has been pointed out that affective/motivational concerns of instructional impact cannot be addressed by using the technique. The instructional objectives and test items measuring their attainment must be aligned or keyed to the instructional content in order to use documented internal review models. Additionally, internal review does not address either adapting instruction to students or the sequencing or structuring of content. For these specific reasons both internal review guidelines and developmental testing guidelines need to be adhered to in order to pronounce the instruction as being valid (accurate — through expert appraisal) and reliable (effective — through empirical testing). The work of Weston (1987a) is seminal to this proposition as she stresses the importance of using both sources of feedback.



Field Testing

Dick (1968) conducted a study that is particularly relevant to data collection and revision activities. In this study, programmed calculus materials were field tested by 85 students from four colleges and universities with four instructors — one from each institution. All of the data generated from this field test were then handed on to the instructional developers (not the original instructors) along with a specified revision heuristic. The revisers chose not to adhere to the heuristic; instead they selected those data sources that they perceived to be the most useful. Frame error rates and subject matter expert review comments were the two most widely used of all of the seven sources for data they had originally been provided. Incorrect responses to frames and student generated comments were occasionally used while test item data, lists of incorrect responses, and a guide relating test items to the content presentation were rarely used. Program revisers were asked if they preferred obtaining program effectiveness data by going through the program personally with three or four students or by obtaining statistical data from 40 or 50 students. Their response indicated that: "There seemed to be a greater acceptance of a more limited type of information from a greater number of students (which appears to provide greater generalizability), and an acknowledgement of difficulty of obtaining suitable guinea pig students" (p. 101). Unfortunately, as the author admits, the study was not taken to its logical conclusion by determining experimentally the effectiveness of the proposed revisions.

There are comparative studies using field tests. These studies will be reviewed later in the section on comparisons of approaches.

Learner Verification and Revision (LVR)

The only difference between developmental testing and learner verification and revision (LVR) is that the latter proposes that the testing of materials be continued throughout the lifespan of the materials or if materials are going to be used by a target population for whom the instruction was not originally designed. For example, if materials are "transported" internationally then extended testing would provide data upon which the instruction might be adopted to this new population of learners (possibly the original instruction was written in a language other than the native language of the new population). Common to both developmental testing and LVR are the proposed phases of one-to-one and/or small group testing.

In one of the few studies employing a revision algorithm, Dupont and Stolovitch (1983) compared expert review with Gropper's (1975) non-validated theoretical revision model which utilizes test subject data. Results from this study investigating the effects of different methods of revision indicated that materials revised using the revision model and test subject data did not produce significantly superior attainment of the objectives when compared to subject matter expert revision using only test subject data. Also, materials revised on the basis of student data produced significantly superior attainment of learning objectives on an achievement test on the manipulation of an SLR camera, but not on a paper and pencil test on the principles underlying the use of an SLR camera when compared to expert (subject matter expert) revision.

For this study, Dupont and Stolovitch (1983) used test subjects considered as being representative of the target population. Ninety junior college students were randomly selected from a population of 200 (a single college). Professional non-teaching staff members of the college served as the four revisers in the study. They had had some experience in evaluating instructional materials. Two of the revisers had more formative evaluation experience than the other two. Each of the four revisers conducted two one-to-one sessions, according to Kandaswamy's (1976) procedures, in order to obtain tutorial LVR data from student test subjects. Reviser #1 applied an adapted version of Gropper's revision model to the data obtained from the first session while the second session, held one

week later, permitted the use of subject matter expert validation only (no formal revision model). Reviser #2 employed the use of subject matter expert validation to the results obtained from the first one-to-one session whereas the adapted version of Gropper's revision model was applied to the data gleaned from the second one-to-one session. Reviser #3 utilized Gropper's revision model for both the first and second one-to-one session data. Reviser #4 exclusively adhered to subject matter validation for both the first and second session data. The two revised versions per reviser (eight versions in all) were randomly assigned to eight randomly selected groups of 10 students.

These authors note that: "to our knowledge, however, Robinson (1972) and Gropper (1975) are the only two who have attempted to develop comprehensive procedures for revising instructional materials based on some explicit theoretical framework" (p. 33) and that:

Robinson's (1972) model is based on taxonomies that were not conceived of in terms of evaluation and revision of instructional materials. The model was designed specifically for evaluating programmed instructional materials (p. 33).

Kandaswamy, Stolovitch, and Thiagarajan (1976) conducted a study to compare the effectiveness of one-to-one and small group-based LVR. One-to-one and group conditions used equal proportions of "high", "medium", and "low" level students. In the one-to-one condition, subjects provided reactions, responses, and remarks to the evaluator/reviser, who did not instruct students when difficulties were encountered but did probe when difficulties were observed. Group-based LVR subjects only provided pre/posttest data and within program response data — no debriefing data were collected. All versions of the instructional material revised under these conditions were administered to groups of students to determine their relative effectiveness. A Scheffe multiple contrast test demonstrated that all revised versions were more effective than the original. There were no significant differences between group-based and one-to-one LVR. A number of evaluators/revisers were utilized in the study and this was a factor in the experimental design. An evaluator/reviser main effect was observed indicating that different evaluators/ revisers produced varying degrees of effectiveness for the materials.

In a similar study, Baghdadi (1980) was interested in determining: if revised materials were better than unrevised materials; if there were any differences between one-to-one and small group LVR approaches; and if different revisers produced varying degrees of effective instructional materials. The instructional material used in the study was a programmed unit designed for graduate students on four different evaluation models. Four evaluators familiar with LVR procedures were hired to revise the materials. Two of these evaluators independently conducted one-to-one evaluations while the other two conducted small group sessions. Then they revised. The two groups of evaluators then exchanged the methods to produce four other versions of the material. The eight revised versions and the original version were randomly assigned to target learners. Revised versions were significantly better in attaining design objectives than the unrevised version. No significant differences were detected between one-to-one and small group approaches. No reviser effect, in terms of producing revised materials of varying effectiveness, was found in this particular study.

A number of researchers have commented on variables considered to be important when designing and conducting LVR sessions. Thiagarajan (1978) discusses how test subjects might be selected. "During group-based LVR, random selection of learners from the target population provides realistic representation" (p. 138). Anecdotally, based on personal experience, the author admits his preference for choosing above-average test subjects for one-to-one testing and then confirm the revisions generated from one-to-one with the results from randomly chosen groups.

Which type of learner makes the best try-out subject has been a topic of much needed discussion. Berthelot (1978) utilized high, medium, and low verbal ability students in an experimental study on tutorial LVR. Although there were no significant differences in revisions, Berthelot obtained best results from the highly verbalizing subject.

Dupont(1977), by inserting feedback questions into an audiovisual training module, and using a group-based approach to LVR, increased the effectiveness of revisions. Gagnon and Stolovitch (1980), utilizing a modified form of feedback questions with third grade children, obtained positive results. They note that for short materials, the tutorial method appears to be more cost effective. The authors present a decision formula that can be utilized to generate a revision feasibility index based on cost-

effectiveness variables. For materials requiring use over a long period time, the group-based approach is recommended as being more feasible.

Debert (1979) has produced a classification scheme for the general kinds of revision prescriptions that instructional developers tend to identify based on the problems encountered by test subjects during LVR materials tryouts. A simple inspection of the categories demonstrates that subjects are able to comment on problems encountered with materials at a level that reflects a broad spectrum of the kinds of decisions that are made in order to produce instruction. However, as Weston (1987a) notes, the revision process can be viewed as mainly subjective or intuitive as there are few guidelines to aid a designer in translating developmental testing data into effective revisions.

A major goal of formative evaluation research has been to produce such validated guidelines. Through the accumulating research literature, there are examples of such efforts. A case in point is the work of Dupont (1980) who has produced an algorithm that facilitates the translation of the data collected into the kinds of revision prescriptions identified by Debert (1979). Stolovitch and Berthelot (1982) have used this approach and reported it as being effective.

The research cited above has been recognized as having political and legal ramifications. Research has indicated that LVR can be used to validate instructional materials. This fact has been used in court in the United States of America in order to lobby for state legislation requiring that instructional materials be subjected to empirical testing. Data from testing would be used to demonstrate the effectiveness of instructional materials before being adopted on a state-wide basis. Komoski and Woodward (1985) note that it has been estimated that as much as 90 to 95 percent of instructional time is spent using some kind of learning material and yet little of this material has undergone empirical testing for quality control purposes. As Vedros (1986) notes: "In research conducted by Komoski from 1967 to 1971, he learned that of the approximate 200,000 then catalogued commercially produced instructional materials on the market, less than one percent had ever been tried out with learners prior to publication" (p. 10). Komoski (1976) in his capacity as director of the Educational Products Information Exchange (EPIE) Institute in New York, a clearinghouse for instructional materials, has been a strong advocate (one amongst many) for the enactment of laws

requiring publishers to demonstrate, using empirical LVR data, that published materials will be effective. Efforts to this end have had some success. Several states, e.g., Florida State Senate Bill 492, Chapter 74 - 337, June 27, 1974, pp. 24 - 25, have legislated learner verification and revision as a requirement for state-adopted instructional materials. However, even when it is required by law to do so, Dick (1986, 1987) provides data indicating that few publishers follow the law to the letter.

Vedros (1986) notes there was a time when publishers had complained that they did not understand what learner verification was nor how to do it. Today, this is certainly not the case. The State of Florida's State Instructional Materials Council branch of the Department of Education (1986) has published strict guidelines that are required by law to be followed for purposes of publishers submission of material being considered for state-wide adoption. The State of Florida's Department of State (1986) has produced a self-instructional module designed to train members of State Instructional Materials Councils on how to conduct a valid evaluation of materials in order that council members would be able to determine whether or not a publisher had adhered to the law.

Komoski's (1976) report on guidelines to be adhered to in order to conduct LVR contains a section on methods for obtaining feedback data:

Techniques for gathering feedback from learners may include, but are not limited to, individual learners and classroom observations, interviews, questionnaires, and criterion-referenced tests. These techniques should be designed to gather data on direct-learning effects and on affective reactions of learners (p. 6).

Komoski (1974) has also discussed the feasibility issue of using just a sample of one test subject in order to collect data for revision purposes.

... it is likely that even one member of the target population may be a useful subject for a verification episode. This sample of one is entirely acceptable

because it is reasonable to assume that any instructional difficulties with a given material that are experienced by one member of the larger target population are very likely to be experienced by other members of that population. Thus, the 'law of parsimony' may perhaps be more elegantly applied in the learner verification and revision process than anywhere else in educational research. This should be remembered by all who are tempted to think that the publisher that carries out a program of large sample learner verification and revision activities necessarily has a better program just because he uses large samples. He may simply be wasting money and time, and passing the cost of that waste on to the schools (p. 379).

Cost effectiveness must be considered as a budget is being considered during the planning phase; the more testing sessions conducted the higher the cost of developing the materials. Engler (1976) has criticized Komoski's position that a sample of one may be sufficient. Engler also notes that perhaps research will not be able to provide strong guidelines about the costs involved in conducting formative evaluations:

There are too many variables to permit a simple generalization about costs. It depends on the subject matter, the grade level, the target population, the media to be used, the instructional strategies employed, and so on (p. 13).

Developmental Testing

In an early and frequently cited study, Robeck (1965) addressed the feasibility of using a single student (one-to-one) to provide the database upon which revision was to be grounded. A revised version was produced based upon the reactions of one test subject (a bright male grade six student) to the prototype. Another revised version was produced based on the reactions of another individual test subject. The three versions of the material (the prototype, version one, version two) were then randomly assigned to three matched groups of students. The two revised versions significantly outperformed the prototype. There was no significant difference between the performance of versions one and two. This study does demonstrate that even revising based solely on a single student can significantly improve instruction. The only problem with adopting this practice is how to select an "appropriate student"?

Baker (1970) was interested in looking at the effectiveness of revised versus original instructional material. A revision heuristic, applicable to error rate data obtained from the developmental testing of programmed instruction, was used in an experiment comparing the effects of first draft materials versus the same materials revised based on the revision heuristic. A mean increment of .68 standard deviation units was obtained from the revised materials over the unrevised versions.

Wager (1983) used Dick and Carey's (1978) procedures for conducting both one-to-one and small group tryouts in an experiment addressing both "the appropriate student" and the number of subjects issues. From pools of high, medium, and low aptitude groups, it was found that high aptitude subjects in the one-to-one condition were more adept at locating inaccuracies and generating instruction for a self-instructional print module when it was perceived as deficient. Low aptitude students identified the more basic problems such as the use of inappropriate vocabulary within the module. A mixed aptitude pool, providing feedback during one-to-one, produced the greatest variety compared to homogeneous aptitude groups. Students who received materials revised based on the feedback from the mixed aptitude group scored significantly higher on the posttest and attained the most

favorable attitude rating than did students who used materials revised either on the basis of feedback from the high or low aptitude group. Wager's study also indicated that materials revised based on feedback from the one-to-one sessions using subjects from the mixed aptitude group were as effective as materials revised based on a combination of one-to-one and small group sessions. The materials revised based on one-to-one sessions from the mixed aptitude group were more effective than materials revised according to feedback from a small group session alone.

During the development of a print self-instructional material designed to train a wide range of professionals interested in producing instructional materials (to the first draft stage), Baker (1972) ran a number of concurrent experiments, all using small group samples, to investigate some of the critical variables involved in the instructional development effort. One study looked at the effect of promoting a revision set on quantity and utility of revision information. Eighteen UCLA volunteer students enrolled in a secondary teacher preparation course were asked to revise first draft versions of text instruction on utilizing the techniques of advance organizers, inspection behavior, and task discrimination. Two conditions prevailed during the experiment. In the treatment group (the "Revision Set" condition), subjects were instructed to read the material and to try to identify specific things that they would change. This was told to the subjects before they went through the materials. In the other condition, control subjects were told, after having gone through the materials, to indicate on the back of the text and on a practice exercise sheet any revision suggestions. This could be conceived as the non-revision set because subjects had not been informed of the revision task prior to beginning the instruction. Two independent scorers rated the utility of suggestions on a scale of one to five and counted the number of discrete suggestions. No significant differences were found between the control groups and the experimental groups. The inter-rater reliability computed was .82 for the number of suggestions and .84 for the utility ratings.

Baker (1972) virtually replicated this study with another experiment involving similar materials. The major difference between the two studies was that the second one employed a professor of the course, instead of using research project staff to administer the materials as was done in the first study. Again, half of the group received the "revision set" while the other half acted as controls. The

number of discrete suggestions and utility of suggestions were also used as dependent measures. Again no statistically significant differences were found between the Revision Set group and the control group. Baker (1972) attempted to interpret the results: "Perhaps the most compelling explanation, applicable to both studies, is that exhorting students to keep revision in mind, without explicitly cueing them to relevant categories of information, is simply not a powerful enough treatment to produce more information" (p. 63). By providing a revision set to test subjects combined with a categorical system for making revision comments, (i.e., directing subjects to attend to the prototype in terms of difficulty, sufficiency of information, clarity, relevance, usefulness, and worth) the developer/evaluator will obtain revision suggestions relating to the categorical system. Anecdotally, the author goes on to state that subjects in both of these studies (in both conditions) did provide valuable revision data relating to both the content and organization of the instruction.

Henderson and Nathenson (1977) were interested in determining whether versions of self-instructional material revised on the basis of student feedback were more effective than original versions. The original materials had been developed and revised using the expert judgement of the design team. The materials were designed for an Open University course (a distance education course) in Elements of Music. The materials comprised all of the instruction students received in the course (students do not attend a formal class — they use the materials at home, submit assignments to a tutor, and sit in a formal examination). Due to the fact that the materials were designed for home use, they were tested by students at home. Developmental testing was comprised of placing questions about the effectiveness of the material in strategic points within the materials. This allowed a test subject to comment on all of the components of the course. A stratified test subject sample of $N=28$ individually worked through the 16 units of the course and made their comments. The materials were then subsequently revised by the design team in light of the data collected. Another group of test subjects ($N=20$) was then selected to use the developmentally tested materials. The two groups were compared, to determine equivalence, on three measures: previous educational experience, musical background, and entry-level pretests. There were no significant differences between the two groups. Data from the first group had been used to revise the material for the second group. The study then compared the two

groups on three dependent variables: performance on posttest, decrease of study time, and/or fewer learning difficulties. The developmentally tested materials significantly out-performed the original version on all three variables.

Comparisons of the Approaches

A number of studies have compared the relative effectiveness of some of the approaches being endorsed in the literature. Newman (1957) was interested in comparing the effectiveness of student versus expert design of studying techniques. A paired associate task of learning the names for each of 20 electrical symbols was used in two experimental conditions. One condition permitted the subjects, all airmen at Lackland Airforce Base, Texas, to study the material using their own set of study techniques. The other condition had instruction structured through the input of five psychologists who were asked to suggest techniques that would be expected to enhance the learning of the paired associates. There were significant differences on a posttest between the two experimental sessions. When subjects were permitted to use their own study techniques, they out-performed subjects placed in a more structured study environment.

Frase, DeGracie, and Poston (1974) conducted a study to investigate the comparability between panel review board results and the results of a pilot test (field test). The adequacy of a competency-based kindergarten curriculum unit was the object of study. The selected jury consisted of five primary school teachers from the lower elementary grades. Each one of the teachers independently rated the quality of the unit through a criterion (quality control) checklist. Ratings were aggregated to provide a global rating across raters. In the other condition, the 20 hour packaged unit was field tested on a group of 12 kindergarten students by a teacher who had developed the unit. The role of this teacher in the study was to independently observe and record pilot test data while another teacher delivered the unit. Evaluation forms and questionnaires were developed to collect a broad range of pilot test data. After a class period, the observer and the deliverer of the instruction each took an hour or so to independently synthesize and evaluate the data. They then incorporated these conclusions into a joint

critique. The evaluative responses of the pilot test team and panel were then compared in order to determine the consistency between the two different sources for evaluation. Particular conclusions from the study indicated that pilot testing was twice as expensive as the panel review. Of critical note was the finding that the pilot testers found fewer problems in the unit than did the review board: "... problems identified by the panel review board were not manifested in 'real classroom problems' as indicated by the pilot test" (p. 35). The authors concluded from this study that empirical evidence indicted panel review was not a suitable substitute for the pilot test.

Golas (1983) presents a study, based on her doctoral dissertation, designed to compare two models of formative evaluation in terms of their ability to produce both instructional and cost effectiveness. One model used student data to ground revisions while the other was an internal review checklist (developed by the author) incorporating instructional editing guidelines. A sample of 10 instructional designers was stratified based on their experience in instructional design and then randomly assigned to one of two groups, the internal review group or the student feedback group. The subjects were 70 soldiers attending a training course and had not previously received instruction on the module utilized in the study. The instructional material consisted of a self-paced, self-instructional print module on the learning task of issuing supplies to customers.

The two groups of instructional designers were given copies of the original materials and told to revise them. One group revised based on Baker's (1974) one-to-one and small group formative evaluation model and the other group based on the internal review guidelines. The two versions of the materials were then compared in terms of their instructional effectiveness and cost effectiveness. There were no significant differences on the instructional effectiveness dependent variable, based on posttest performance, between the two experimental conditions. Striking differences in terms of cost effectiveness were identified in the study: "... the cost of developing modules in which student data was collected was substantially greater than the cost of developing modules when the editing guidelines were used." (p. 18)

Generalizations Based on the Literature Review

A major generalization can be made based on the literature reviewed in this dissertation: revised instruction is more effective than the original unrevised instruction. Many of the reviewed studies have concluded this and there are others to support the generalization. Case studies by Markle (1967); Clarke (1975); Repicky and Hotvedt (1977); Wile (1979); and Weston (1987b) indicate that instruction revised based on using student feedback is more effective than unrevised original instruction. Another generalization to be made is that most authors and researchers recommend using internal/expert review in combination with test subject feedback. The exclusive use of either approach is not frequently recommended although authors taking this position have been discussed.

What is not clear from the literature related to formative evaluation are the particular guidelines related to the design and conduct of testing sessions for purposes of obtaining feedback useful for revision. Wiener (1954) has defined feedback:

Feedback is a method of controlling a system by reinserting into it the results of its past performance. If these results are merely used as numerical data for the criticism of the system and its regulation, we have the simple feedback of the control engineers. If, however, the information which proceeds backward from the performance, is able to change the general method and pattern of performance, we have a process which may well be called learning. (p. 61)

This definition emphasizes utilizing feedback in an attempt to control entropy and establish order within a system. Levin (1974) emphasizes that feedback plays an important role for accountability within education:

For operational purposes, I will define a perfectly accountable system as one that maximizes the utility or satisfaction of the relevant constituencies

... Essentially, an accountability system is a closed loop reflecting a chain of responses [by an organization] to perceived needs or demands [of a constituency]; ... a set of activities that emerges to fill those demands; outcomes that result from those activities; and feedback on outcomes to the [constituency]. The feedback may generate new demands or a regeneration of old ones . . . ; a new or an altered set of activities may . . . be produced; and the loop is completed again with feedback to the [constituency]. The result is that an accountability system is continuous and dynamic when the linkages are tight and information is generated and transmitted freely. (p. 375)

It is clear that developmental testing data could be used to hold a developer accountable to the original design objectives of the product. Similarly, a publisher could use developmental testing data in order to account to educational institutions desiring to purchase effective and efficient learning material; such data could be cited as evidence.

Prototype research and development activities occur in a number of different areas and educational researchers have used a number of metaphors from these areas to help in conceptualizing the rationales for conducting prototype testing. Markle (1967) has made an analogy between developmental testing and the testing of experimental aircraft in flight engineering:

No airplane manufacturer would build a hundred airplanes straight from the initial drafting board and send them aloft to see if 90 percent of them would fly. Wind tunnels and stress tests save pilots. The test pilot takes risks to save others. The conditions under which the test flights are run may not represent the total sample of real-life tasks the airplane must bear up under, but the fact that it has passed the initial tests augurs well for later success. The man who runs an airline, like the teacher in the classroom, cannot afford to try-out unknown products on the job, although the latter often does. (Time

rather than lives are lost.) Developmental testing, in industry and education, saves the consumer, at considerable cost to the producer, from faulty products. (p.125)

During prototype development, testing is geared to gathering information on the failures, not on the successes, with this information permitting the tester to investigate the cause of failure, as Englemann (1983) has noted. McCrory (1964) has likened the process of test and revise during instructional development to the scientific method of hypothesis testing. Popper (1959; 1973) has argued that virtually all problems can be tackled by using the same methodology. First a problem situation is identified and then a tentative solution to the problem situation is formulated. Errors can be eliminated by empirically testing the tentative solution.

Baker (1973) draws upon research and development approaches used in pharmacology to create a metaphorical link with the developmental testing of instructional products. In either of the two disciplines, the toxicity of the product can only be assessed by empirical testing. Unintended side effects (possibly toxic/detrimental to a learner) and interactions between the treatment and particular target population characteristics can only be evaluated on a Post Hoc basis. Jacobs (1962a, 1962b) describes the technical considerations and procedures underlying the process of validating and demonstrating reliability of criterion tests. Student test subjects' responses are used for purposes of item analysis. The author makes an analogy between test development and instructional programming, using members of the target population as sources for validation purposes.

Although there is general agreement as to why prototype testing should be done, there is a lack of consensus as to how it should be done. Markle (1967) refers to developmental testing as a "debugging operation" emphasizing the quality control dimension and characterizes it by its clinical approach: "Data tend to be "clinical" in the sense that they are the result of close observation of, and often interaction with, a few students, singly or in small groups, who may or may not be representative of the target population" (p. 111). In one-to-one testing, the second test subject serves as a check on the problems identified, explicated, and resolved through the feedback generated by the first subject. The purpose of this phase is to produce a workable instructional design product whereas :

Large group testing undoubtedly produces unwanted suppressions of individual candid reactions of the sort that occur when a friendly observer accepts any and all statements of the student as data worth consideration. In a group, a student who would like to ask a "stupid" question may not ask for it, fearing that others might laugh. The 'stupid' question may be about some key confusion that the programmer failed to spot. (Markle 1967, p. 111)

Baker and Alkin (1973) agree with Markle's analysis:

The two factors of exposure to untried instruction and the need for rich data combine to suggest that the initial development tests be conducted on a relatively small number of students and the data collection procedures be designed to provide the completeness of data required. Smaller samples represent a loss in the reliability of the data and a shortening of turn-around time for analysis and revision. Since initial development often has an exploratory character, it would be unwise to interrupt the development process with lengthy intermissions caused by analysis of too much data. Thus, the solution of smaller samples, more intensively analyzed, would appear to be an appropriate one to the data collection dilemma at this stage of development (p. 410).

Baker and Alkin (1973) have indicated that developmental testing usually requires more than one trial. Dick (1977), a proponent of the three-stage model for conducting formative evaluation, suggests that it is important to build in at least one iteration of each of the three phases: one-to-one, small group, and field test. Although he admits that it is possible that evaluation should never cease, he suggests that local conditions will impact on terminating evaluation activities. Variables such as time constraints, resources at hand, and pre-set criterion levels of performance could all affect the

formative effort. Dick (1980) summarizes his position on the differences between one-to-one and small group testing procedures:

Most writers refer to either two or three stages of formative evaluation (Dick, 1977; Markle, 1967; Baker, 1974). Some refer to an early, prototype stage and a later operational development stage; some, to one-on-one, small-group, and field testing stages. In either case, the procedures and techniques are basically the same. An early version of a product is tried out with a few learners in an almost clinical setting. Obvious problems are detected, changes made, and further try-outs conducted. As the development progresses, large groups of learners and more realistic learning situations are employed. The final stage of testing in the field involves the next-to-final version of the product, and administrative feasibility of its use is determined. Final changes are made, and the product is released for use in the field (p. 3).

Dick (1980) recognizes that the first stage, the one-to-one phase, can be considered to be the most crucial phase in terms of generating data to modify the instruction. Scott and Yelon (1969) agree that the use of an individual student as a test subject has the compelling rationale of saving a lot of time, money, and effort through identifying problems at an early stage of development. Most authors suggest that one test subject is not sufficient even though research has indicated that materials revised based on the exclusive feedback from one subject can be effective. Dick (1980) recommends that the issue of validity of revisions based on just one subject: "... might be addressed on a larger scale in combination with the issue of student profiles to provide more data and a better understanding of the usefulness of individual learner data" (p. 5).

Then there is the issue of how to go about selecting test subjects for inclusion into developmental testing sessions. Dick (1980) notes that a common selection criterion for one-to-one testing is to use an above average, average, and below average student. With this kind of a recommendation it becomes

clear that the number of test subjects to be utilized interacts with characteristics of the sample. On the other hand Englemann (1983) opts for poor test learners. Higher level students, he argues, will provide such a low error rate that little information will be gained. He de-emphasizes the common recommendation to use a representative sample:

The goal of data in a tryout is to help identify flaws in a product, not to discover a population tendency. To identify flaws we use the type of tools that are implied by quality control. We inspect instruction through direct observations. We also augment this effort with data-collection tools. We respond to observed faults in instruction by providing quick responses, either in the form of replacements for parts of the program, or training (p. 16).

Nathenson and Henderson's (1977) position is to stratify test subjects on the basis of entry behavior as is evidenced by performance on pretest measures. The underlying rationale behind this recommendation is that test results are often used as an indicator of effective instruction. Relationships between pre and posttest scores can be used to generate profiles of different kinds of test subjects providing formative feedback. Dick (1979) has been a proponent of building such profiles:

Since that first phase, usually referred to as the one-to-one stage, is often the most critical in terms of modifying the instructional approach, it is important to know how the participants in the one-to-one are selected. A common technique has been to select above average, average and below average students to go through the instruction on an individual basis with the formative evaluator. It would appear that identifying these learners on the basis of specific entry skills and knowledge would provide a sounder basis for comparisons of the error profiles and attitudes which they generate. Such comparisons could be made to determine the usefulness of that data in the revision of the instruction. For example, comparing the effectiveness

of instruction revised on the basis of the data from these three different types of 'operationally defined' students could indicate which, if any, of the types provides consistently better data for instructional revision purposes. (p. 3)

Also at issue are the roles that participants should play in order to obtain feedback data. Markle (1967) proposes a highly interactive developmental testing session:

The tester does more than take notes; he interprets or filters what he observes. A great deal of the significant behavior of the student is covert - he is thinking, evaluating, and so forth, to himself. The ideal tester would seem to be one who is both sensitive to subtle indications of student problems and a master at solving them, at spotting gaps in the presentation, unpredicted confusions, over-generalizations from previous examples or failures to generalize. His job is to observe the student, to let the student teach him, rather than to defend his own ideas of how the student should be responding to the brilliantly designed instruction (p. 124).

The roles that participants play during developmental testing will affect the kinds of data being obtained. If a test subject is asked to go through the material as best they can and complete the tests, then errors made by the test subject will provide the basis for an item analysis. But Nathenson and Henderson (1977) have recognized that error data inherently has limitations: "... observation that comparison of pre- and post-test performance, though capable of identifying weak programme components, is inherently incapable of revealing reasons for weakness in a way which permits the generation of solutions." (p. 9) It is for this reason that most authors recommend gathering additional data. Horn (1964) suggests that the following kinds of data may be obtained from a student tryout:

1. identification of objectives not attained by the test subject

2. identification of test items giving the test subject problems and determination as to why the items do so
3. identification of test subject errors
4. provides the test subject the opportunity to qualitatively judge the material

Baker and Alkin (1973) have identified several kinds of data considered as being important to collect:

"The three types of data important to determining product effectiveness in the prototype or initial development stage are: 1) student performance of post-instructional tasks; 2) student attitudes and satisfaction with program; and 3) unanticipated side effects engendered by the program" (p. 409).

These authors have noted that the following can be considered as sources of data useful in revising instruction:

1 Expert judgement

2. Program-response data — error patterns on practical exercises

3. Program-critique data (obtained by):

- a) Learners can circle difficult words, write comments, or place question marks in the materials.
- b) Single learners can be probed as they proceed through the materials.
- c) A post-instructional diagnostic questionnaire where students can suggest revisions or via interview.

A much neglected component of both the formative evaluation prescription and research literatures is the debriefing procedure. Thiagarajan (1986) provides a brief sketch of what kinds of data can be expected to be gathered through the use of structured interviews. Stolovitch (1982) notes that this procedure is very useful with test subjects who have difficulties expressing themselves in writing. There is a paucity of research investigating the effects of debriefing data or on how debriefing data might compare with other possible methods of data collection and resulting revisions. Authors advocating one-to-one testing do not appear to emphasize debriefing as much as authors who advocate small group testing. Generally, those who do endorse the debriefing procedure usually are not explicit

In explicating how a debriefing session is to be conducted. Gels, Burt, and Weston (1984) have noted that there are defined guidelines for conducting interviews with clients in social work agencies and for patients in psychological clinics. If professionals in other areas can be trained to conduct these kinds of interviews then there seems to be little reason why similar training can not be conducted for debriefing. Fear (1973) has written a textbook on conducting evaluation interviews and Warwick and Lininger (1975) offer guidelines for conducting survey interviews to which Stolovitch (1982) suggests can be adapted to evaluation. To-date, though, it remains unclear as to how data gathered from debriefing would compare with other kinds of data gathered during developmental testing.

Of final note is the issue of using developmental testing procedures with materials designed using a medium other than print-based self-instructional modules. Although medium is not a concern to this study it is still an important consideration in designing instruction. Hartley and Burnhill (1977) advocate that textbooks be formatively evaluated. These authors provide a large number of guidelines, based on research to a great extent, for the use of prose, tables, graphs, and typographical design in the production of instructional text. In spite of the plethora of guidelines, many addressing rather complex typographical concerns, the authors advocate the use of empirical testing of the text material

There is one overall guide-line which is applicable to *all* instructional text.

This is that initial versions need to be tried out with samples of the target population for whom they are intended and revised on the basis of results obtained.

This should be done before the material is finally committed to publication. One cannot assume just because one has, for example, constructed a graph, that the point being made is now automatically understood by the learner. The effectiveness of a graph - and anything else - needs to be tested (p. 71).

Sedlik (1971) has collated a large mass of research from a number of disciplines on the obtaining and utilization of feedback for the improvement of mediated instructional products. The focus is on the collection of data from the target audience during and after film production. Gropper, Lumsdaine, and

Shipman (1961) and VanderMeer, Thorne, and Montgomery (1961) have conducted a number of studies using learner feedback for the purposes of improving mediated instruction, e.g., films, television programs, filmstrips, etc. Learner feedback tended to be comprised of response data to criterion posttest items. Typically these studies show that revising based on this kind of data significantly improves performance on posttests for versions of instruction using this type of feedback data. More recently, Baggailey (1986) has reported on the use and effectiveness of the Program Evaluation Analysis Computer (PEAC) system to obtain audience response data for the purposes of formatively evaluating television programmes. Although formative evaluation of mediated materials or textbooks may prove to be more challenging than self-instructional printed materials, research has indicated that empirical testing has improved these materials.

Research Questions

Dick (1980) has commented on the kinds of research activities that need to be undertaken:

The state-of-the-art in formative evaluation at the present reflects a relatively mechanistic approach to the process . . . What is needed is a more theoretical approach which would a) suggest the kinds of data which should be collected and b) suggest how the data should be interpreted by the designer and used to revise the instruction. Such an approach would indicate the extent to which the instruction reflects the theory and is instructionally effective and the ways in which the instruction should be improved to better reflect the theory (pp. 5 -6).

Burkholder (1982), reviewing the literature since 1960, views the research related to the revision of instructional materials in three particular areas:

1. revision methodology

2. case studies explaining the revisions made; and
3. the derivation of principles for revising instruction

Within the area of revision methodology, emphasis has been on comparing the effectiveness of instruction revised based on either one-to-one sessions, small group sessions, or on both one-to-one and small group sessions. This research strategy appears to address only one of Baker and Alkin's (1973) research questions related to formative evaluation: "1. What are the major types of data to obtain and the major sources of such data? and 2. What is the impact of evaluation data upon the modification of programs?" (p. 393). The effectiveness of revised versions of instructional material under varying experimental conditions is a frequent dependent variable within the developmental testing of instructional materials research literature because the reason, in practice, for conducting testing sessions is to obtain data for purposes of revision. There is a methodological problem that has plagued this line of research. The effectiveness of revised materials studies contaminate the study of effects of testing variables since the reviser or the revision heuristic is a key variable. The way in which the evaluator / reviser interprets the data gathered from testing sessions will impact on the kinds of revisions that are made. Since heuristics or algorithms for implementing revisions are not presently agreed upon, it appears that the effectiveness of revised instructional material will remain strongly related to the subjective ability of the evaluator / reviser to remedy problems identified through developmental testing. There is a need to compare the kinds of raw data generated from small group and one-to-one testing before it is pronounced that the two methods produce comparable results. Possible differences in the data gathered during one-to-one and small group testing will remain obscured if only student posttest scores are used to index and study the effectiveness of various revision methodologies. Even when research indicates that one methodological approach produces higher posttest performance than another or even when no differences are cited, it remains unclear as to how the revised materials differ qualitatively in terms of their revised design because researchers never describe these outcome differences or similarities. Similarly, researchers have tended not to describe the differences and/or similarities in the raw data that was used in order to revise the instructional product.

Although research, to this point, has tended to emphasize the relative effectiveness of revisions, the importance of the sources and kinds of data produced from them should not be ignored. It is quite feasible in supposing that the data produced under one-to-one and small group conditions will differ, otherwise conducting evaluations using a combination of the two approaches would be unnecessary. It is important to describe the kinds of data obtained from both one-to-one and small group testing before these data are used in order to prescribe revisions. Not until this kind of descriptive research is undertaken can it be concluded that that one-to-one and small group are comparable methodologies.

The literature reviewed recommends that, within one-to-one/small group sessions, evaluators utilize a number of different sources for collecting feedback data. Proposed sources of feedback data include: student test subject articulations of problems while trying to learn from the material, student test subjects suggestions for resolving these encountered problems, posttest errors, affective reactions to a questionnaire, and the debriefing of test subjects after the test session. It can be hypothesized that the assignment of differing roles that the participants could play during a testing session (e.g., an actively intervening and probing evaluator versus a more passive observer) might affect the data from these sources. However, no research until now has systematically described the data from these different sources of feedback under such conditions. A better understanding of these data utilizing various formative evaluation methodologies should ultimately facilitate our understanding of the revision process — but it is not the intent of this research to look at the process of revision nor the effectiveness of revision although there may be implications for revision.

Both the reviews of the instructional design textbook literature and the research literature have revealed a great deal of variability in prescribed guidelines for the design and conduct of developmental testing sessions. Because of this variability, it was decided that it would be fruitful to determine which guidelines were perceived by "expert" practitioners of instructional design as being the most useful. A survey of expert practitioners' actual practices was conducted in order to identify these guidelines. Once these guidelines were identified, the amount of agreement between proposed procedures from the literature and procedures practiced by "expert" instructional designers in the field could be determined. The results of a survey of a large number of instructional designers will be discussed in

the next chapter.

O

O

CHAPTER III

QUESTIONNAIRE ADMINISTRATION

In this chapter the results of a large survey of instructional designers will be presented. The questionnaire was designed to solicit the procedures that practitioners actually use in order to design and conduct developmental testing sessions for purposes of revising prototypes of instructional material. Technical considerations for the development, mailing, and analysis of the results of the questionnaire will be discussed.

Rationale for the Questionnaire

The literature review reveals that the guidelines for developmental testing are incomplete. An author describing how to design and conduct testing sessions will not necessarily address all of the eight variables revealed by the content analysis of rules and guidelines conducted for the purpose of this study. A given author's recommendations will often conflict/contradict with another author's set of guidelines. The research literature reviewed reveals that sometimes significant differences, in terms of effective revision, are found in favor of one-to-one testing. Sometimes differences have been found in favor of small group testing. Other times research has revealed no significant differences between the two approaches. None of the research reviewed directly investigated the effects that different participant roles might have on the data gathering process, (the amount and kind of data might be expected to vary with different roles that the players could assume).

In consideration of the above discussion, it is difficult to conceive how practitioners make informed decisions based on a literature that offers so many diverse recommendations. Possibly practitioners and publishers might be overwhelmed by the lack of consensus on how to developmentally test to the point where developmental testing is not taken seriously as an important step in systematic instructional design. Presently there is no justifiable way to decide which guidelines from amongst the

many will be effective and efficient in generating data for the purposes of revision. It is because of the need to develop a consistent set of guidelines that the following research was conducted. By surveying instructional designers in the field, the extent to which practitioners developmentally test and the extent to which the various techniques for obtaining feedback data are utilized can be described. The results of the questionnaire will permit a comparison to be made between recommendations from the literature of developmental testing and the practice of developmental testing.

Designing the Questionnaire

As a result of reviewing the textbooks on systematic instructional design and the related literature of formative evaluation (focusing on the developmental testing and revision of prototype instructional products), a set of recommendations for conducting developmental test sessions was compiled. These recommendations represent 14 textbooks and one dissertation devoting substantial effort to the explication of conducting developmental testing and revision activities. Table 1 reflects this review of the textbook literature. Table 1 is a synthesis of the results of the content analysis which has been previously discussed. It represents a matrix of recommended guidelines by author along eight presumed critical variables for developmental testing. In order to reach consensus with regards to the synthesis of the recommended guidelines for designing and conducting developmental testing sessions, two independent judges reviewed the 15 references paying particular attention to the sections in the texts dealing with formative evaluation, i.e., developmental testing. Notes were made about the procedures. Then the two judges compared notes resulting in a consensus of the recommended guidelines. The results of this consensus are represented through in the individual cell entries of Table 1.

The content analysis of the various procedures (rule statements) provided the basis for the construction of a classification system for all of the procedures being endorsed by the authors sampled. This permitted the "mapping" of the variability of recommendations between authors and highlighted the conflicting nature of the procedures taken as a whole. It also revealed that not all authors present a

Table 1
Categorization of Guidelines Proposed by Selected Authors for Designing and Conducting Developmental Testing

79

	Criterion Performance	Procedures/ Number of Students	Population Characteristics	Role of Student	Role of Developer	Data	Basis of Revision	Testing Environment
Abedor (1971)	80%/80% on posttest; 80% + satisfaction on the survey	small group (N = 9 - 12); debriefing	abilities (varying; SAT scores)	learner; commentator; revisor	administrator; observer; prober	pre/post tests; attitudinal survey; student comments; tape recordings of sessions & debriefings; developer observations & notes	error rates (pre/post); survey ratings; comments	cites Horn
Baker & Schutz (1977)	test until performance variability eliminated	1 to 1 over an extended period of time; later product tryouts to simulate field test	-	learner	developer may be different from administrator; observer	pre/posttests; developer observations & notes	error rates (pre/post and embedded items) comments	correspond to natural conditions
Bell & Abedor (1977)	80% + of students correct on test item; 70% + satisfaction on the survey	small group (N = 4 - 6); debriefing	representative; ability (varying)	(see ABEDOR)	(see ABEDOR)	(see ABEDOR)	(see ABEDOR)	-
Bjerstedt (1972)	predicted or expected competencies	1 to 1; small group	representative; traits (not too limid)	learner; commentator; think aloud	administrator; observer; tutor	pre/post tests; survey; student comments; observations	error rates (pre/ post and embedded items)	informal; quiet room; lots of time
Davis, Alexander & Yelon (1974)	achievement of objectives	small group; debriefing	representative	learner; commentator	administrator; prober	pre/post tests; survey; student comments	error rates (pre/posttests); comments	-

Table 1
 Categorization of Guidelines Proposed by Selected Authors for Designing and Conducting Developmental Testing

	Criterion Performance	Procedures/ Number of Students	Population Characteristics	Role of Student	Role of Developer	Data	Basis of Revision	Testing Environment
Deterline & Lenn (1972)	achievement of objectives (all students all objectives)	1 to 1 (N = 3); debriefing	representative	learner; commentator	administrator; tutor; revisor	pre/posttests; embedded items; student comments; tape recordings of sessions & debriefings	error rates (pre/post and embedded); comments	-
Dick & Carey (1978)	-	1 to 1 (N = 2+); small group (N = 10 - 20); debriefing	representative	learner; commentator	administrator; observer; prober; tutor; revisor	pre/posttests; attitudinal survey; student comments; developer observation and notes	error rates (pre/post and embedded items); survey ratings; comments; observations	-
Friesen (1973)	posttest performance; embedded items (less than 7% error rate); achievement of objectives	1 to 1; small group (N = 6); simulated field test (N = 30)	representative	learner; commentator	administrator; observer; tutor	pre/posttests; embedded items; student comments	error rates (pre/post and embedded items); comments; observations	-
Gropper (1975)	-	1 to 1; groups; debriefing	representative	learner; commentator	administrator; prober; tutor occasionally	pre/post tests; survey; student comments	error rates (pre/posttests); comments	-
Horn (1966)	predicted or expected competencies	1 to 1 (N = 3) for novice developers; small group (N = 3 - 4) for expert developers; debriefing; re-test after revision	representative	learner; commentator	administrator; observer; prober; tutor; revisor	pre/posttests; attitudinal survey; student comments; developer observation and notes; time	error rates (pre/post and embedded items); survey ratings; comments; confidence level	Informal; quiet room

Table 1
Categorization of Guidelines Proposed by Selected Authors for Designing and Conducting Developmental Testing

	Criterion Performance	Procedures/ Number of Students	Population Characteristics	Role of Student	Role of Developer	Data	Basis of Revision	Testing Environment
Johnson & Johnson (1975)	posttest performance; embedded items; achievement of objectives; satisfaction on the survey	1 to 1 or small group (N= 3 - 6); debriefing; re-test after revision	pretryout with anyone; representative	learner; commentator; revisor	administrator; observer; prober; revisor	post test; embedded items; attitudinal survey; comments	error rates (pre/post and embedded); survey ratings; comments	-
Lawson (1974)	predicted or expected competencies	1 to 1 (a few); small group	representative; ability (below average)	learner; commentator	administrator; observer; prober	pre/posttests; comments; attitudinal survey; developer observations & notes	error rates (pre/post and embedded items); observations; survey ratings	-
Markle (1978)	posttest performance (acceptable)	1 to 1	traits (highly motivated & verbal)	learner; commentator; think aloud	prober; tutor; revisor	pre/posttests; embedded items; comments	error rates (pre/post and embedded items) comments	-
Merrill & Tennyson (1977)	posttest performance; confidence ratings	1 to 1 (N=2); re-test after revision; small group (N = 15)	a student	learner; commentator	administrator; observer; tutor	pre/posttests; comments; developer observations & notes	error rates (pre/post); confidence level; comments; observations	-

Table 1
Categorization of Guidelines Proposed by Selected Authors for Designing and Conducting Developmental Testing

	Criterion Performance	Procedures/ Number of Students	Population Characteristics	Role of Student	Role of Developer	Data	Basis of Revision	Testing Environment
Thiagarajan, Semmel, & Semmel (1974)	achievement of objectives	1 to 1 (N = 4 - 5): qualitative testing; re-test after revision; simulated field test: quantitative testing	traits (verbal & enthusiastic)	learner; commentator; think aloud	administrator; observer; prober; reviser	comments; embedded items; developer observations & notes; prerequisite, pre/ posttest and survey; (from simulated field test only)	comments; observations; error rates (pre/ post and embedded items); survey ratings	-

complete set of guidelines. All of the authors, in differing ways, focused their attention to explain the following distinct variables:

1. The criteria for a validated product.
2. The number of test subjects (learners) required during developmental testing and the general testing procedures.
3. The characteristics of the test subject population.
4. The role of the test subject population during developmental testing and revision.
5. The role of the instructional designer/evaluator during developmental testing and revision.
6. The data.
7. The basis for revision.
8. The testing environment.

These categories, then, represent presumed critical variables involved in the developmental testing and revision process. The isolation of the presumed critical variables facilitated the construction of the first draft of a questionnaire designed to solicit respondents' agreement with the various rule statements found in the literature. In addition, respondents were to be asked to provide a justification for their chosen procedures. This was included because the literature review revealed a paucity of justifications or rationales for many of the procedures authors were advocating. The first draft version of the questionnaire required respondents to:

- provide a general definition of developmental testing and revision.
- characterize the test subject population.
- explain how test subjects were differentiated with regards to entry skills, I.Q., etc.
- indicate how many test subjects were typically used.
- elaborate on the role of the test subject during developmental testing and revision.
- explicate the role of the instructional developer during testing and revision.

- indicate when testing is terminated.
- give some indication as to when revision is terminated.
- explain what data are obtained during developmental testing and which procedures provided the most useful data for revision.
- indicate how data gathered from testing are analyzed.
- provide a basis upon which instruction is revised.
- indicate how explicit and teachable respondents' procedures are for testing and revision. Respondents were also asked to provide documentation, if existing, for these procedures.
- provide demographic data of respondents' professional activities, recent publication record, and how they were trained as instructional designers.

An open-ended section was included to indicate general reaction to the survey and possible ways to improve it.

The content of the questionnaire reflected the rules, guidelines, heuristics, and algorithms for the design and conduct of developmental testing sessions for purposes of obtaining data for revision as they had been stated in the literature of instructional design. The options for each multiple-choice item were keyed to specific recommendations. Consequently the questionnaire items dealing with rules and procedures selected, often verbatim, the array of recommendations in this literature. Technical content validity and reliability considerations were addressed through pilot testing the questionnaire.

Pilot Testing

As a pilot test eleven "experts" in the area of instructional design were sent a first draft of the questionnaire and asked to respond to it while noting comments, questions, and suggestions on format and content. A follow-up 10-minute telephone interview with each respondent was conducted approximately 10 days after its mailing. All of their comments were compiled and common trends in the suggested revisions were identified. The experts who participated in the pilot testing of the first draft questionnaire were selected based on their expressed interest in the research, development, and application of systematic formative evaluation methods to instructional products. All of the the eleven

experts had either written a textbook or journal article related to the subject matter contained within the draft questionnaire, written or supervised dissertation work related to the topic, acted as independent consultants on major instructional design projects, or supervised instructional designers in a business, military, or academic environment. The experts who participated in the pilot test were judged to be representative of the target population to be sampled in that they were engaged in similar kinds of employment and employment responsibilities and had expressed a professional commitment to formative evaluation. As a result of their feedback, the following revisions were made:

- a cover letter for the final draft of the questionnaire added a general definition of both developmental testing and revision. Respondents were directed to an additional section in the survey if they could not accept the general definitions; this allowed respondents to explain their unique procedures. Another section was added to the questionnaire to enable respondents to describe how they would test and revise if not operating under practical constraints, e.g., time and budget, to provide a comparison of respondents' typical procedures with what they felt would be ideal.
- sections soliciting demographic information were deleted because it was found that readily accessible demographics on the sample were available from the National Society for Performance and Instruction's Membership Directory
- another response set was added to permit respondents to sequence their testing and revision activities on a time line.
- a question was added which attempted to identify whether or not designers have found that successive test subjects generally replicate each others' comments.
- questions which originally included responses to both testing and revision were divided in two, one for testing and one for revision.
- as an "incentive" to prompt respondents to fill in and return the questionnaire, a section was added allowing the respondents to indicate if they wished to receive a description of the development of the questionnaire, results of the questionnaire, or both.

The final version of the questionnaire and its accompanying cover letter focused on asking respondents (practitioners in the field) to select the recommended practices to which they in fact adhered. Each item on the questionnaire, with a few exceptions, addressed some aspect of developmental testing (e.g., characteristics of test subjects). The items contained rule statements which represented the range of prescriptions in the selected literature. Respondents were asked to select those statements which reflected their practice. Additional items concerned the utility of developmental testing, the procedures

for revision, and the teachability and documentation of the respondents' testing and revision procedures.

The 18 page, 18 item final version of the questionnaire (see Appendix C) required respondents to:

- characterize the test subject population.
- explain how test subjects were differentiated with regards to entry skills, I.Q., etc.
- indicate how many test subjects were typically used.
- elaborate on the role of the test subject and the designer during developmental testing and revision.
- report when testing and revision was terminated.
- explain what data were obtained during developmental testing and which procedures provided the most useful data for revision.
- indicate how collected data were analyzed, to provide a basis upon which instruction was revised.
- indicate how explicit and teachable the respondents' procedures were for testing and revision.

Respondents were also asked to provide available documentation for these procedures.

For all questions posed, respondents could utilize an "OTHER" response to identify procedures that the questionnaire did not address. After selecting/identifying their procedures, respondents were asked to provide a rationale, in open-ended format, for their utilization of the specified procedures. All respondents were also requested to identify, in open-ended format, their "ideal" procedures, (i.e., as if there were no timeline or budget constraints operating).

Sampling Procedures

The sample receiving the questionnaire were experts in the area of systematic instructional design. The National Society for Performance and Instruction, a leading professional organization for training specialists, was chosen as the source from which to build a pool of possible respondents. All of NSPI's presenters from the 1984 annual meeting were surveyed. A Participant Directory of the meeting included the names and addresses, job descriptions, research interests, and other demographics

of the presenters at the conference. In addition it suggested that a great many of the presenters would be able to provide valid responses to the questionnaire because of their expressed interest in using systematic instructional design procedures. Each presenter was sent a questionnaire. Presenters who had already participated in the initial pilot testing of the questionnaire or whose textbooks had been taken into account for the compilation of the questionnaire were dropped from the sample.

Mailing Procedures

The final version of the questionnaire and its accompanying cover letter (see Appendix C) focused on asking respondents (practitioners in the field) to select the recommended practices to which they in fact adhered.

On July 5, 1984, 206 participants were sent a questionnaire. Mailing procedures did not attempt to provide anonymity for the sample. On September 11, 1984, 115 non-respondents were sent out a reminder card (see Appendix C). Table 2 reports the final response rate for the various categories of the sample. Some of the sample returned an uncompleted questionnaire indicating they were not involved in instructional design. Some reported that they used a completely different approach that did not permit them to fill out the questionnaire. Part of the sample did not return the questionnaire.

Table 2 has been constructed so that those who in some way responded can be compared to the total sample (N=206). There was a total response rate to the questionnaire of 62%. Table 3 presents a demographic breakdown of the samples' principal type of employment both within and across sample response-type categories, (i.e., total sample; respondents, non-respondents, and those not involved in instructional design). Table 4 presents a similar demographic breakdown for the samples' most significant employment responsibility. For both of these important demographic variables, there were no practical differences across sample response-type categories. This suggests that the respondents used in the data analysis (N=90) were probably representative of the total sample (N=206).

Table 2.

Questionnaire Respondents and Non-Respondents

	<u>TOTAL</u> <u>SAMPLE</u> N=206	<u>%</u>	<u>RESPONDENTS</u> N=127	<u>%</u>
Respondents used in the data analysis	N=90	44%	N=90	71%
Respondents not involved in Instructional Design	N=32	16%	N=32	25%
Respondents using another approach	N=5	2%	N=5	4%
Non-respondents	N=79	38%	-	-

Total Response Rate= $127/206 = 62\%$

Table 3.

Demographic Breakdown of Samples' Principal Type of Employment

	Total	Respondents Used in the Data Analysis	Respondents Not Involved in I.D.	Non-respondents
	N=206	N=90	N=32	N=79
	%	%	%	%
Academic/Educational Institution	29%	29%	25%	27%
Consulting/Contracting Firm	20%	21%	25%	22%
Independent Consultant	4%	6%	6%	3%
Private Business/ Industry	19%	20%	9%	4%
Government Agency (Federal, State, City) Other Than Military	7%	8%	9%	18%
Military (As Civilian or Service Personnel)	7%	4%	3%	15%
Utility Company (Such as Power, Communications, etc.)	7%	9%	6%	11%
Medical Facility	4%	2%	6%	-
Other	2%	1%	9%	-
Total	99%	100%	98%	100%

Note that some columns do not total 100% due to rounding.

Table 4.

Demographic Breakdown of Samples' Most Significant Employment Responsibility

	<u>Respondents</u> <u>Used in the</u>		<u>Respondents</u> <u>Not Involved in I.D.</u>	<u>Non-respondents</u>
	<u>Total</u>	<u>Data Analysis</u>		
	N=206	N=90	N=32	N=79
	%	%	%	%
Training Director/ Manager	26%	29%	16%	27%
Developer of Training and/or Non-training Solutions	21%	23%	6%	22%
Trainer (Delivery/ Training)	1%	-	-	3%
Developer and Trainer (Equal Responsibilities)	7%	9%	6%	5%
Researcher, Consultant	22%	22%	31%	18%
Teacher/Professor	14%	12%	16%	15%
Other	8%	4%	16%	11%
Total	99%	99%	101%	101%

Note that columns do not total 100% due to rounding.

Post Hoc Statistical Hypotheses Generation

The intent of the questionnaire was to obtain data in order to describe the practices of expert instructional designers working in the field. For this reason descriptive research questions were generated based on the categorical system containing the eight variables presumed to be important in designing and conducting developmental testing sessions. Specific research questions had been established *a priori* — they are reflected in the first 18 items of the questionnaire. On a *post hoc* basis, individual statistical hypotheses were generated on a by individual question (or item) basis. At times it was necessary to generate several statistical hypotheses for particular questions. This was decided after raw frequencies had been calculated on all of the returned and usable questionnaires. These hypotheses were oriented towards the quantification of use of particular guidelines. Sometimes raw frequencies and their corresponding percentages were sufficient to address a particular research question. When this was the case (high frequencies indicating overwhelming adherence to particular rule statements) statistical hypotheses were not generated as frequency data was enough to describe the samples' response. When this was not the case, due to considerable variability in the calculated frequencies, *post hoc* statistical hypotheses were generated in order to test for statistically significant differences between proportions of responses.

Data Analysis Procedures

Two types of data emerged from the questionnaire. For open-ended response sets, (i.e., the "ideal" procedures) a content analysis was performed. The objective data coded from the questionnaire was analyzed using the SPSSX Multiple Response procedure as respondents (N=90) were permitted to select more than one response per question. Non-parametric tests of significance (Chi-square), when appropriate, were used to reveal significant differences between proportions of responses in order to test various research hypotheses of interest.

Results

Results will be presented by item according to the sequencing of questions within the questionnaire. When a result for a particular statistical test of significance is presented, the null hypothesis is first stated. Following a null hypothesis statement is the result of a *post hoc* Chi-square non-parametric statistical test. Just before that result is given, the specific *a priori* research question is presented. There are occasions when a number of research questions are aligned to an individual questionnaire item (question). There are occasions when only raw data will be reported which appropriately address a specific research question. The raw frequency data used in the data analysis of the questionnaire results can be found in the questionnaire located in Appendix C. The frequency of each response is written beside each one of the boxes that respondents used to indicate their selections.

Questionnaire Item #1:

H_0 . In developmentally testing prototype instructional materials, there is no difference in the proportions of respondents using test subjects in order to gather data versus those respondents using a subject matter expert, or a peer, or some other review mechanism to obtain information as to the effectiveness of the instructional design.

Result: During the developmental testing of prototype instructional materials, a significant proportion of the respondents indicated that they used test subjects (as opposed to some other review mechanism), $\chi^2_{calc} = 65.94$, for $df = 1$, $\chi^2_{0.05} = 7.88$, reject H_0 . Eight respondents did indicate that instead of using test subjects, subject matter or peer reviews were used. This small percentage of the sample then went on to qualify this particular response by stating that these other review mechanisms were utilized only when test subjects were unavailable or when timelines pre-empted the use of test subjects.

H_0 : Of those respondents endorsing the use of test subjects, there is no difference between the proportions of respondents who only use test subject feedback data and those who also obtain information from SME's or peer reviews as to the effectiveness of the materials.

Result: A significant proportion of the respondents indicated that in addition to using test subjects, subject matter experts or peer reviews are also used, $\chi^2_{\text{calc}} = 9.24$, for $df = 1$, $\chi^2_{.005} = 7.88$, reject H_0 .

Item #2:

H_0 : Of those respondents who endorse using test subjects, there is no difference in the proportions of respondents who revise during the testing session, after the testing session, or both during and after the testing session

Result: A significant proportion of the respondents indicated that as a general strategy, revision takes place after the testing session, $\chi^2_{\text{calc}} = 46.0183$, for $df = 2$, $\chi^2_{.005} = 10.6$, reject H_0 .

H_0 : There is no difference in the proportions of respondents who rely solely on criterion test data, on test subjects' comments solely, or on both test data and test subjects' comments.

Result: A significant proportion of the respondents indicated that as a general strategy, revision is based on both the test data and the comments made by test subjects, $\chi^2_{\text{calc}} = 29.9126$, for $df = 2$, $\chi^2_{.005} = 10.6$, reject H_0 .

H_0 : There is no difference in the proportions of respondents administering an attitudinal survey for purposes of revising versus respondents not administering a survey.

Result: A significant proportion of the respondents indicated that they did not base revision on an attitudinal survey, $\chi^2_{\text{calc}} = 26.22$, for $df = 1$, $\chi^2_{.005} = 7.88$, reject H_0 .

Item #3:

H_0 : There is no difference in the proportions of respondents using only high level subjects, high and low level subjects, enthusiastic and highly verbal subjects, or a mix of high, medium, and low level subjects (a representative sample of the target population), or some other requirement for test subject selection.

Result: A significant proportion of the respondents indicated that test subjects, utilized during the developmental testing and revision phase, in some way represented high, medium, and low level/aptitude performers, $\chi^2_{\text{calc}} = 123.74$, for $df = 4$, $\chi^2_{.005} = 14.86$, reject H_0 .

Item #4:

H_0 : There is no difference in the proportions of respondents differentiating test subject level/ability based on data generated for developmental testing (e.g., a pretest), pre-existing available data, or some other differentiating mechanism.

Result: A significant proportion of the respondents indicated that in order to differentiate test subject level/aptitude before the test session, pre-existing data (e.g., a standardized test, past achievement data, job performance measures, etc.) were utilized as opposed to having the developer generate this kind of information (e.g., pretest results), $\chi^2_{\text{calc}} = 53.75$, for $df = 2$, $\chi^2_{.005} = 10.6$, reject H_0 .

Item #5:

H_0 : There is no difference in the proportions of respondents utilizing only small group testing procedures or both one-to-one and small group testing procedures.

Result: There is no significant difference in the proportion of respondents using solely small group sessions as opposed to both one-to-one and small group sessions, $\chi^2_{\text{calc}} = .16$, for $df = 1$, $\chi^2_{.05} = 3.84$, accept H_0 . Eight respondents indicated that they only used one-to-one sessions.

Research Question: HOW MANY ONE-TO-ONE AND/OR SMALL GROUP SESSIONS ARE TYPICALLY USED?
HOW MANY TEST SUBJECTS ARE ASSIGNED TO A GROUP? Result: When asked to provide the actual

number of sessions and test subjects, respondents tended to respond by providing ranges, e.g., 2-3 test subjects. The mean responses using the minimum and maximum values of the ranges are reported:

a) number of one-to-one sessions: min = 3.2 / max = 4.6

b) number of small group sessions: min = 2.1 / max = 3.4

c) number of test subjects typically used in each small group session:

min = 7.1 / max = 10.8

Item #6:

Research Question: WHAT IS THE ROLE OF THE TEST SUBJECT DURING DEVELOPMENTAL TESTING?

Result: A large majority (81/90) of the respondents place the test subject into an active role during developmental testing. This role permits test subjects to freely comment on problems with the instructional materials.

Item #7:

Research Question: WHAT IS THE ROLE OF THE TEST SUBJECT DURING REVISION?

Result: Based on raw frequencies, respondents perceived the role of the test subject to provide pre- and posttest data as well as to comment, either verbally or in writing, on the problems they encounter within the instructional materials.

Item #8:

H_0 : Respondents do not perceive successive test subjects as providing new (i.e., non-redundant) information.

Result: A significant proportion of the respondents thought that different test subjects often provide qualitatively different information during successive developmental testing and revision sessions, $\chi^2_{\text{calc}} = 15.5236$, for $df = 1$, $\chi^2_{.005} = 10.6$, reject H_0 .

Note that items 9, 10, and 11 refer to the following broad research questions: HOW INTERACTIVE IS THE DEVELOPMENTAL TEST SESSION?; WHAT IS THE NATURE OF INTERACTION?

Item #9:

Research Question: WHAT IS THE ROLE OF THE TEST SUBJECT DURING DEVELOPMENTAL TESTING?

Result: Raw frequencies reveal that the role of a test subject during a developmental test session is to ask questions or make comments. It is the developer's role to encourage the test subject to do so. Only 3% of the responses indicated that the test subject is not permitted to make comments and ask questions.

Result - Raw frequencies reveal that 87.3% of the responses indicated that the test subject is unrestricted in the amount of interaction that can be initiated with the developer, in terms of asking questions and making comments to the developer.

Item #10:

Research Question: WHAT IS THE ROLE OF THE DEVELOPER DURING DEVELOPMENTAL TESTING?

Result: By far the majority of the respondents, 86.7%, place the developer into an active role during developmental testing. This would logically follow from previous questions placing the test subject into an active role as well. The two roles would naturally complement one another. The most frequent response indicated that if a problem arises, the developer asks "why?", pinpoints the problem, and gives just enough help to get the test subject back on track.

Item #11:

Research Question: WHAT IS THE ROLE OF THE DEVELOPER DURING REVISION?

Result: The two most frequent responses recommend that a developer debrief a small group of test subjects with the agenda for the debriefing arising from criterion test data, attitudinal reactions, and elicited test subject comments (i.e., from a comment sheet). Revision is to be based from these sources. Since earlier in the questionnaire respondents indicated that attitudinal surveys were not

conducted, respondents must secure attitudinal data from other sources such as observational and inferential techniques — or some less formal methodology.

Item #12:

Research Question: WHEN DOES DEVELOPMENTAL TESTING STOP?

Result: Frequencies indicate that respondents to the questionnaire terminated developmental testing activities after some criterion response pattern was being met (pre-set criteria or otherwise).

Naturally, project deadline, client, money available, etc., were the over-riding factors.

Item #13:

Research Question: WHEN DOES REVISION STOP?

Result: Frequencies indicate that the instructional designers responding to this survey stop revising when some consistent criteria are being met, e.g., 80/80, 90/90, or after several sessions revealing consistent results.

Item #14:

Research Question: WHAT ARE THE SOURCES OF DATA FOR REVISION?

Result: When directly asked: "How do you obtain your data during developmental testing?" respondents most frequently indicated:

- from criterion pre- and posttests.
- by recording information (by note-taking, audio-tape, or video tape).
- by debriefing test subjects after the test session.
- from test subjects' notations written on the instructional materials.
- from spontaneous verbal comments that the test subjects have made during the test session.

Within the same question, the sample was asked to indicate from a listing of the procedures involved in data collection activities (the list arising from the literature of instructional design), the most useful of all of the procedures. Respondents most frequently selected three from the eleven choices. They are listed in rank order:

1. By debriefing test subjects (after the session)
2. From criterion pre- and posttests
3. From test subject notations written on the instructional materials

Item #15:

Research Question: HOW IS THE DATA GATHERED FROM DEVELOPMENTAL TESTING SESSIONS ANALYZED?

Result: Frequencies indicate that the major factors affecting the extent of data analysis procedures are the resources available (e.g., budget and time constraints). Developers debrief test subjects after their completion of the material, and then tally test results, attitudinal data, and test subject generated comments. The test subject and the developer cooperatively interpret the data. Developers then test out the revisions generated from a previous session on the next test subject(s). Strong emphasis is placed on error analysis to find out why individual test subjects make errors. To a lesser extent, developers compare pre- and posttest data with a pre-set standard.

Item #16:

Research Question: WHAT IS THE BASIS FOR REVISION?

Result: After having obtained their data, 34% of the sample indicated that revision was based solely on "obvious", self-evident weaknesses within the instructional materials. Only 11% of the responses indicated that they consistently followed a prescribed algorithm or heuristic for revision. Six percent of the responses indicated that the revision process was intuitively based.

Item #17A:

Research Question: ARE DEVELOPMENTAL TESTING PROCEDURES TEACHABLE?

Result: When queried as to how explicit and teachable are procedures for conducting developmental testing sessions, the respondents indicated that their procedures are well documented in the literature or in handouts. They stated that their procedures could be/are taught to others. Another large group of the respondents said that even though their procedures were not documented - they could easily be documented. By analyzing the frequency data from this question, it appears that 72% of the total number of responses perceive that developmental testing procedures are teachable.

Item #17B:

Research Question: ARE REVISION PROCEDURES TEACHABLE?

Result: When asked the exact same Question as 17A but for revision, their response pattern was virtually identical to the results from 17A. From an analysis of the frequency data for this question, it appears that 68% of the total number of responses perceived that revision procedures are teachable.

Item #18:

After having responded to the multiple response formatted questions, respondents were requested to describe ideal procedures for developmental testing and revision procedures. As many of the members of the sampled population worked in business and industry environments, it was conceivable that working constraints (e.g., timelines, budget, client prerogatives, and availability of test subjects) may not have permitted these instructional designers to utilize what they might have considered to be ideal ways to collect data for the purpose of revising the instruction. Question 18 in the questionnaire asked the sample to consider, if possible, the variables noted in the questionnaire and to outline, using a free response format (one blank page was provided), their conception of optimal testing procedures for revision purposes. The ideal procedures reported could then be compared with both the literature's recommendations and the actual procedures that the respondents selected within the questionnaire for

the purposes of identifying discrepancies and concurrences both within and between these three contexts.

All of the responses to question 18 were collated in order to facilitate a content analysis of the free response data. A frequency count of those respondents who did respond to the Question revealed that only 45% could be considered "usable". Some respondents, for example, just provided text or journal citations as representing the ideal. Respondents in this category were dropped from the pool of usable responses. A closer inspection of the data revealed that generally respondents did not closely adhere to the variables as outlined in the questionnaire and that there was considerable diversity in the ways that the question was addressed. Some respondents provided very general strategies for conducting developmental testing and revision sessions while others were very specific. Some respondents elaborated more on some variables than others although all were considered to be important. Because of the low response rate and the general lack of adherence in addressing the identified variables, a formal content analysis was not conducted.

An informal examination of responses suggests some major themes or trends. In general, the reported ideal responses did parallel the literature's recommendations, (i.e., novel or different procedures did not emerge). The following responses, selected directly from the raw data, were chosen to reflect the range of ideal procedures that instructional designers advocate:

"My experience has been in the area of computer-based education for use in the industrial setting. Budgets and schedules are always tight but we have basically followed the procedures outlined in most of the literature. Most of our testing would be considered more informal than those outlined in the literature but we have found it to be very functional and we actually needed very little revision in the field test cycle. In a perfect environment, we would like to do one-to-one testing, revise and then do at least a couple of test cycles and revise. However, that takes people away from their work and the company loses productivity. So a practical approach works best for our needs."

"I have filled out the form as a composite picture of the various formative evaluations we do in projects at company name. Since they vary so much (ranging from computer-based technical skills to video-supported management education) it is obviously difficult to say much of a standardized nature. In the training business, time/budget constraints tend to dictate methodology more than anything else. We expect our developers to be creative and use whatever techniques work given real world constraints."

"I consider the specification of 'ideal' testing/developmental conditions a counter-productive effort. It gives developers a way out of having to develop the wide range of methods required to deal with the wide range of constraints faced in the 'real world'. I would rather see us all face up to the fact that the constraints we encounter range from minimal to monumental and that if we are to consider ourselves 'master' practitioners, we'd better be able to vary our practice accordingly."

So few respondents attended to the request that they sequence their testing and revision activities on a time line that analyses were not conducted. Respondents had also been requested to provide a rationale, in open-ended format, for their utilization of their selected developmental testing and revision strategies. Again, so few respondents completed these sections within the questionnaire that analyses were not conducted.

Figure 6 presents a summary of the procedures which practitioners endorse for designing and conducting developmental testing sessions.

Chapter IV provides a comparison of the results of the data analysis of the questionnaire with the recommended guidelines from the literature. A profile of expert procedures, based on the data analyses is presented and discussed.

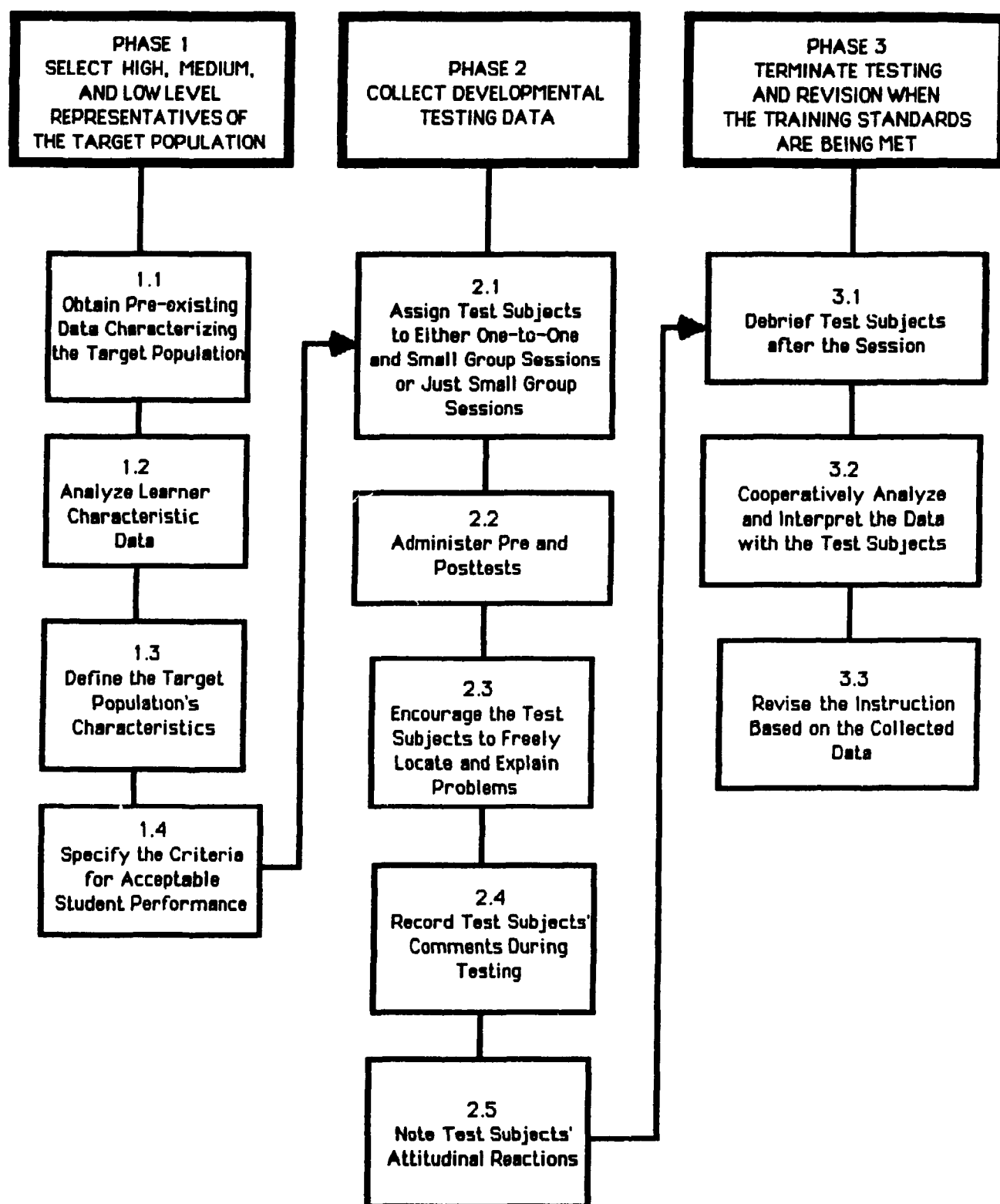


Figure 6. Procedures for the Design and Conduct of Developmental Testing

CHAPTER IV

A COMPARISON OF RECOMMENDED AND ACTUAL DEVELOPMENTAL TESTING PRACTICES

The results of the data analysis will be reported in terms of a comparison between the recommendations from the literature and the reported practices of the sampled practitioners. The recommendation from the literature will be presented first and then the reported practice (questionnaire finding). The two will be compared. Note that the use of the term significant, as it relates to findings, refers to tests reported in Chapter III. Based on the results of the data analysis, a composite profile of expert procedures for the design and conduct of developmental testing sessions will be presented and discussed. Research questions which emerge based on the data analysis of the questionnaire will be discussed.

Discrepancies Between Recommended and Actual Practices

Recommendation: It is virtually universally suggested that test subjects (learners) be used in the developmental testing of prototype instructional materials.

Finding: All respondents to the questionnaire indicated that they used test subjects. Eight respondents did indicate that instead of using test subjects, subject matter experts or peer reviews served instead. These respondents then went on to qualify this particular response by stating that these other review mechanisms were utilized only when test subjects were unavailable or when timelines pre-empted the use of test subjects. This finding strongly concurs with the literature's recommendation to use test subjects.

Recommendation: The literature consistently endorses the use of peer reviews in addition to using test subjects.

Finding: A significant proportion of respondents endorsed this procedure as well.

Recommendation: Some authors recommend revising during the testing session; some after the session; and others suggest revising at any time. There is no strong consensus.

Finding: A significant proportion of the practitioners indicated that as a general strategy, revision takes place after the testing session.

Recommendation: Most authors indicate that revisions should be based on both criterion test data and test subjects' comments.

Finding: A significant proportion of the respondents concurred with the recommendation from the literature. The samples' responses, taken as a whole, does indicate that some practitioners place more emphasis on criterion test results than test subject comments and vice versa. Three kinds of approaches to evaluation emerge: quantitative, qualitative, and both quantitative and qualitative.

15.9% of the respondents said that they relied exclusively on criterion test data. Focussing only on test subjects' comments were 26.6% of the respondents. The 58.5% of the respondents who utilize both criterion test data and students' comments might be termed "eclectic" in their procedures.

Recommendation: The literature strongly suggests collecting attitudinal data, usually by means of a formal attitudinal survey instrument.

Finding: A significant proportion of the respondents do not administer an attitudinal survey. This result sharply conflicts with the recommendations in the literature.

Recommendation: Most authors advocate that the test subjects used during developmental testing should be characterized as being representative of the target population for whom the instruction was designed. This usually means a mix of high, medium, and low level/aptitude subjects.

Finding: A significant proportion of the survey sample agreed that test subjects, in some way, should represent high, medium, and low level/aptitude performers.

Recommendation: Related to the above recommendation, authors diverge on suggestions for differentiating test subject level/ability. Some recommend pretest scores, others standardized aptitude tests, and still others past achievement data, etc. The literature is by no means clear as to what the characteristics of test subjects should optimally be nor how to select individuals.

Finding: A significant proportion of the respondents indicated that in order to differentiate test subject level/ability before the session, pre-existing data (e.g., a standardized test, past achievement data, job performance measures, etc.) were utilized as opposed to having the instructional designer/evaluator generate this kind of information (e.g., pretest results).

Recommendation: There is extreme variability among authors' suggestions for the kind of sessions and the number of subjects to be utilized during developmental testing. Some recommend conducting clinical one-to-one sessions only; some recommend small group sessions; and some recommend both one-to-one and small group. Authors also vary as to the number of one-to-one sessions to be conducted; the number of small groups; as well as to the number of test subjects within a small group. In general the literature asserts that several test subjects be used and/or several types of sessions. This implies that there is some variability in test subject generated data.

Finding: The surveyed practitioners' responses indicate that in practice there is no significant difference in the proportions using solely small group sessions as opposed to both one-to-one and small group sessions. Only eight respondents indicated that they used just one-to-one data. When directly asked to provide the actual number of sessions and test subjects for both one-to-one and small group conditions, the practitioners tended to respond by providing ranges, e.g., two — three test subjects. The mean responses using the minimum and maximum values of the ranges are reported below in Table 5.

Table 5.

The Mean Number of Test Subjects and Sessions

	<u>Minimum</u>	<u>Maximum</u>
Number of One-to-One Sessions:	3.2	4.6
Number of Small Group Sessions:	2.1	3.4
Number of test subjects typically used in each Small Group Session:	7.1	10.8

Respondents were asked if test subjects, during successive developmental testing and revision sessions, a) generally provided different kinds of information, b) different subjects generally replicated each other, or c) both of these conditions had been observed depending on the context. A significant proportion indicated that different test subjects often provide qualitatively different information during successive sessions. This could indicate a rationale for using several test subjects and more than one type of session. By varying the number of test subjects and sessions, much of the variability of test subject generated data supposedly can be identified and addressed.

Recommendation: Already noted is the qualitative/quantitative distinction with regard to the kinds of data that can be collected that the literature offers. Numerous suggestions on ways to obtain these kinds of data include:

- scores from criterion pretest and posttest
- responses to an attitudinal survey
- responses to the probing (during the test session) of test subjects
- recorded test subject behavior (by note-taking, audio tape, or video tape)
- data from debriefing test subjects (after the test session)
- test subject notations written on the instructional materials
- test subject comments placed on a revision comment sheet

- notes that were taken on examples, illustrations, information added, and changes to sequences made while the test subject was tutored through the instructional material
- spontaneous verbal comments that the test subjects have made during the session

Although by no means consistent in making recommendations (as can be seen from the above suggestions), instructional design textbook authors generally recommend a mix of both qualitative and quantitative data collection strategies. It appears that most would recommend collecting pre- and posttest data, attitudinal data, test subject generated comments while going through the material, designer/evaluator observational data, and debriefing (interview) data.

Finding: When asked directly, "How do you obtain your data during developmental testing?", respondents most frequently indicated:

- from criterion pre- and posttests
- by recording information (by note-taking, audio tape, or video tape)
- by debriefing test subjects after the session
- from test subject notations written on the instructional materials
- from spontaneous verbal comments that the test subjects have made during the test session

Finding: When asked to indicate the most useful of all of the varying procedures, respondents chose three strategies. These three strategies were then rank ordered by their frequency of perceived utility as revision data sources. In rank order, they are:

1. By debriefing test subjects (after the session)
2. From criterion pre- and posttests
3. From test subject notations written on the instructional materials

Recommendation: As previously mentioned, the kinds of data that an evaluator decides to collect will affect the roles that both the evaluator and test subject will play during both developmental testing and revision. In general, a majority of the authors suggest that both evaluator and test subject be

encouraged to play an active role during the testing session. Further interaction is facilitated by debriefing after the session — the data can be discussed in an interview situation with the results of the testing sessions determining the agenda. Those who take a strictly quantitative position place the test subject in a more restrictive role — subjects learn the material as best they can and then take the test.

Finding: A large majority (81/90) of the respondents place the test subject into an active role during developmental testing. This role permits the test subjects to freely comment on problems with the instructional materials. Raw frequencies reveal that the role of a test subject during testing is to ask questions or make comments. It is the designer/evaluator's role to encourage the test subject to do so. The majority of respondents, 86.7%, place the designer/evaluator into an active role during developmental testing. This would logically follow the active role of the test subject; the two roles would naturally complement one another. The most frequent response indicated that if a problem arises within the instructional materials, the designer/evaluator asks "why", pinpoints the problem, and gives just enough help to get the test subject back on track. Raw frequencies indicate that 87.3% of the responses to the questionnaire do not restrict the amount of interaction that can be initiated with the evaluator, in terms of asking questions and making comments to the evaluator.

Recommendation: Most authors place the student into a passive role as far as revision is concerned although a very few authors do suggest that test subjects can be asked to actually make the revisions. Most authors suggest that it is the test subject's behavior that will lead an evaluator into making revisions.

Finding: Based on raw frequencies, practitioners agreed with the most frequent recommendations. Test subjects are not actually revising the materials; they indicate problems to the designer/evaluator. This suggests that, as frequently recommended in the literature, test subjects are providing a data base on which the evaluator can ground revisions.

Recommendation: Most authors suggest revising "appropriately" based on the data that have been collected and analyzed. Beyond this, very few authors relate the test data to particular revision actions,

(i.e., provide heuristics or algorithms for revision indicating that certain kinds of errors or problems dictate certain kinds of revisions).

Finding: A major factor affecting the extent of data analysis, for revision purposes, is the resources available (e.g., budget and time constraints). It appears that designers debrief test subjects after their completion of the material, and then tally test results, attitudinal data, and test subject generated comments. The test subject and the designer cooperatively interpret the data for the purposes of identifying problems within the material. Designers then test out the revisions generated from a previous session on the next test subject(s). Strong emphasis is placed on error analysis to find out why individual test subjects make errors. To a lesser extent, designers are comparing pre- and posttest data with a pre-set standard. After having obtained and analyzed their data, 34% of the practitioners indicated that revision was based solely on "obvious", self-evident weaknesses within the instructional materials. Only 11% of the responses indicated that they consistently followed a prescribed algorithm or heuristic for revision. Six percent of the responses indicated that revision was intuitively based. It is interesting to note that the roles that designer and test subject play during the testing session fundamentally affect the revision process. For example, if the test subject is only permitted to learn the material and take the test, revision must be based exclusively on the test data.

Recommendation: Authors are vague in prescribing when to terminate developmental testing and revision procedures. Some suggest using performance criteria such as 80/80 (80% of the learners score 80% or higher on test materials). Most authors do not offer a precise criterion.

Finding: Respondents indicated termination of developmental testing was based upon some criterion response pattern being met (pre-set or otherwise). Project deadline, client, money available, etc., were the over-riding factors. Similarly for the termination of revision, the instructional designers responding to the questionnaire indicated that they stopped revising when some consistent criteria were being met, e.g., 80/80, 90/90/, or after several sessions revealed consistent results.

Finally, respondents were asked if both developmental testing and revision procedures were teachable. The response pattern was strikingly similar for the two activities. A large majority of

respondents either stated that their procedures could be/are taught to others or they stated that even though their procedures were not documented — they could easily be documented and therefore were teachable.

A Profile of Expert Procedures

Based on the response patterns that emerged from the questionnaire, the following represents a "profile" of developmental testing and revision procedures:

1. Select a "representative" sample of test subjects for whom the instruction has been designed.
2. Either conduct several one-to-one sessions and then conduct small group sessions or just conduct small group sessions.
3. Pretest the subjects, administer the learning material, interact with the subjects whenever a problem is noticed in order to locate and define the problem, give just enough help to get the subjects back on track, note attitudinal reactions of the subjects, posttest the subjects, and finally debrief the subjects.
4. Cooperatively analyze and interpret the data with the test subjects.
5. Revise, after the test subject(s) have been debriefed, based on test subjects' behavior during the session.
6. Test out your revisions during the next session.

Discussion

It should be noted that the above profile is a composite; it does not necessarily represent any one practitioner. It does dramatically parallel a general set of recommendations prescribed in the literature; practitioners generally seem to follow developmental testing prescriptions to a great extent.

The first procedure requires that a suitable sample be drawn from the target population. In order to address individual differences within the target population it is suggested to select test subjects who reflect the range of talent within the target population. If the target population is heterogeneous in nature then the test subject sample should be stratified by "high, medium, and low" levels of ability. This will insure that possibly different kinds of concerns raised between test subjects from the three ability levels will be addressed through developmental testing. Revisions would naturally reflect these different kinds of concerns. If the target population is characterized by a restricted range of talent, (i.e., a homogeneous target population) then stratifying by ability is inappropriate due to the lack of variability. In this case, test subjects might be selected randomly from the target population. If there are other variables which characterize the target population, e.g. age, sex, socio-economic status, etc., and these variables are considered to be important in defining the target population (a major component within instructional design models is to define the target population) then test subjects can be stratified along these dimensions as well. It should be noted that a developmental tester is aware of who the target population is because the results of an analysis of the target population has been utilized in order to create the instruction being tested. This knowledge should be used as input into the test subject selection process.

The second procedure in the profile remains ambiguous. It will be more cost-effective if one-to-one sessions could be eliminated by utilizing only small group sessions. Developer/evaluator time spent on developmental testing would be reduced and less test subject time would be required. In a commercial or industrial training environment, this would reduce the amount of time that test subjects would be away from the job. However, one-to-one sessions might be able to contribute unique data unattainable in small group situations. In general the literature recommends that several test subjects be used during both one-to-one and small group settings. With regards to the number of test sessions to be employed, it is generally asserted that sessions be repeated as often as is required with different test subjects until: a) the learning criteria are reached and b) an appropriate sample of the target population has been covered. Related to the issue of the number of test subjects and the number of sessions required during testing is the environment in which testing is occurring. Self-instructional

print-based materials are often studied at home on an individual basis. Henderson and Nathenson (1977) and Nathenson and Henderson (1979) have insisted that developmental testing should take place in an environment very similar to the actual environment in which the material will eventually be used. If self-instructional materials are intended for use on an individual basis (as is the case with distance education coursework) then they must be developmentally tested on that basis in order to insure that valid results are obtained. This line of argument does have much appeal and could be used to challenge some of the empirical results that have attempted to determine which testing strategy is more effective (one-to-one versus small group). The argument does not apply to materials that are not designed to be used on a self-instructional basis. One of the motives for conducting developmental testing strategy research is to determine which strategy is the most cost-effective and efficient. According to Henderson and Nathenson's position, with regards to the testing environment, questions addressing cost effectiveness should not be directed at eliminating the one-to-one phase. Henderson and Nathenson have proposed that even the presence of the developer/evaluator during the one-to-one session disrupts the testing environment — students taking distance education courses do not have the opportunity to have developers/evaluators sit with them on a one-to-one basis. For this reason it has been proposed that embedded evaluation/feedback questions be placed within the prototype materials in order to overcome the contaminating presence/influence of a developer/evaluator. Test subjects are debriefed at a later time by the developer/evaluator. The only problem with an embedded question approach to developmental testing is if the embedded questions are eliminated in the final version. Then, of course, the testing environment remains different from the real user environment.

The third procedure also requires clarification. All authors advise utilizing a written or performance posttest. This will facilitate the quantification of results from developmental testing. In addition to criterion test data, most authors recommend collecting qualitative data as well. This finding parallels the literature of program evaluation where researchers are suggesting that the distinction between quantitative and qualitative methods is not helpful. Cook and Reichardt (1979) and Guba and Lincoln (1983) propose that a compromise between the two methodologies is necessary. Guba and Lincoln (1983) recommend employing both quantitative and qualitative methods due to the inherent

strengths and weaknesses of the two basic approaches. They state that: "Each approach has advantages: quantitative methods have greater precision and are mathematically manipulable, while qualitative methods are richer and can deal with phenomena not easily translatable into numbers" (p. 323). Dick and Carey (1985) have recognized the difficulty of summarizing test subjects' qualitative comments generated from developmental testing sessions:

It is almost impossible to summarize these comments in tabular or graphic form. It is better to try to relate each of these comments back to the instructional materials themselves, or to the objective in the materials to which they refer" (p. 228).

Related to the distinction between quantitative and qualitative kinds of data, are the roles of the participants in testing sessions. Clinical one-to-one testing affords an opportunity for the participants to interact and produce qualitative data such as comments and suggestions about the learning materials. Small group testing does not allow for the same kind of clinical, probing activities as one-to-one. Small group sessions do not easily allow for intensive probing and are often viewed as initial approximations to field testing. For these reasons they tend to involve the test subjects in a less interactive role. This is probably why most authors recommend a combination of the two testing strategies — both kinds of data can be collected. The debriefing procedure commonly associated with small group testing does permit for intense probing, within a group problem-solving environment, in order to supplement the quantitative data gathered from group testing. Many authors suggest a post-instructional debriefing during which the developer / evaluator sets the agenda (based on data obtained during developmental testing), probes for difficulties perceived by the test subjects and summarizes the group's comments. This could explain why the roles of the participants (test subjects and the developer/designer) are likely to differ in one-to-one as compared to small group sessions as well as during testing as distinguished from debriefings following testing sessions. During a developmental testing session (excluding debriefing), most authors reviewed suggest an intermediate role for test

subjects within the continuum of "test subject-as-traditional student" (a passive role) and "test subject-as critic" (an active role). This intermediate role encourages test subjects to make occasional comments, often in response to limited questioning by the developer/designer. Practitioners, on the other hand, tend not to limit in any way the amount of interaction occurring between the test subjects and the developer/evaluator. Practitioners place far more emphasis on the utility of debriefing data than the literature suggests.

The purpose of developmental testing is to obtain data about what needs to be revised. However, various kinds of information can be generated, e.g., criterion test data, test subject comments, and observations of non-verbal behavior, etc. Essentially two different kinds of data can be collected: qualitative and/or quantitative. In all cases the purpose is to obtain data for the revision of the instructional product. Some authors in the literature emphasize relying on quantitative test data while others suggest relying on both quantitative and qualitative feedback. Decisions about the kinds of data to be obtained obviously will influence the ways in which the sessions are conducted, (i.e., affect the assigned roles of the designer/evaluator and the test subjects).

One unique source for data that can be used to obtain both quantitative and qualitative feedback is a formally administered attitudinal survey instrument. Likert-type scaled items can be quantified and open-ended (free response format) questions can provide qualitative data. The third procedure in the profile recommends noting the attitudinal reactions of the test subjects. Most of the textbook authors in the literature recommend administering an attitudinal survey in order to tap affective reactions to the material. Brandenburg (1982) presents a heuristic for evaluation questionnaire design and management that applies to getting the kinds of affective data that instructional design textbook authors say is important during prototype testing. He presents a hierarchical classification system for identifying different levels of information inherent within any evaluation questionnaire. Brandenburg (1982) argues, based on his own research, that most evaluation questionnaires do not have the impact that was intended because training professionals either lack skills in constructing evaluation questionnaires or do not consider these skills to be important.

A statistically significant proportion of the respondents to the questionnaire indicated that they do not administer an attitudinal survey. It is quite possible that Brandenburg's data is relevant to this finding. It is also possible that practitioners are using less formal techniques such as observational and/or inferential techniques in order to attend to attitudinal reactions to the prototype. Respondents did indicate that attitudes were attended to although not through a formal questionnaire.

The fourth and fifth procedures from the profile deal with revision activities. There is very little agreement between authors as to what "revise appropriately" really means because of the tremendous variability of revision procedures. In fact many of the procedures conflict across authors. Most authors place the student in a passive role as far as revision is concerned. Few authors suggest that test subjects be asked to actually make the revisions; most suggest that the evaluator make revisions based on the database generated from the testing sessions. Quantitative and qualitative student data may indicate when there is a problem and where it occurs. More complicated is the determination of the cause. Probing the test subject further may assist the developer/evaluator in this determination, but the developer/evaluator is ultimately responsible for it.

There are two kinds of data, which emerge from developmental testing sessions and debriefings, upon which revision can be based: a) outcome (quantitative) and b) process (qualitative). Quantitative data is usually obtained from posttests. The result from using a heuristic for the design of systematic instruction should be an alignment of instructional objectives, instructional materials, and the tests. Consequently, diagnosis of weaknesses in the instructional materials involves backtracking from test items which produce sizeable error rates to those portions of the instructional materials which are responsible for the errors. Assuming that the designer has followed the heuristics, revision based on quantitative data should be relatively straightforward if the test items are valid. With qualitative data, given the amount and kind of such data, it is not surprising that there is a paucity of developed procedures which might be applied to qualitative data such as protocols gathered during developmental testing or debriefing sessions. It is implied that redundant test subjects' comments and suggestions indicate inherent problems in the materials. Such comments may also be used to confirm the posttest data — test subjects' negative comments on particular segments of the instructional materials should

correlate with errors on posttest items keyed to that content. Caution must be used in interpreting criterion test results when a developer/evaluator assumes a highly active role during testing to the point where the developer/evaluator overrides the instructional materials and revises on-the-spot based on "tutoring" the test subject through areas of encountered difficulty. This can contaminate the posttest results because they will no longer be a reflection of just the materials; posttest results will be a reflection of both the materials and the "tutoring". Recording the whole session by some means should facilitate the analysis and interpretation of this kind of qualitative data.

From the above discussion, it appears that procedures for revision are more clearly described for the use of criterion test data than for qualitative data. Authors and practitioners generally emphasize a standard based on achievement on a posttest that will guide revision. The designer/evaluator should continue to revise and test until the student test subjects attain the objectives of the instructional materials. Possibly for this reason, the final procedure in the profile requires that revisions be tested during a follow-up session.

The results of the data analysis of the questionnaire when compared with the literature's recommendations strongly suggests that practitioners are unable to conduct as many iterative cycles of test-revise-test as the literature recommends. Respondents to the questionnaire noted that they would like to be able to follow the literature's recommendation but they identified constraints such as budgets and deadlines as factors which produce constraints. Practitioners rarely use an algorithm or heuristic for revision even though they exist in the literature. From the practitioners' point of view, it appears unreasonable to expect instructional designers to follow such revision prescriptions as presently exist because revision "guidelines" are for the most part general and vague and because of real world resource constraints.

Conclusions

From the procedures for developmental testing, recommended by the authors reviewed and adhered to by the practitioners sampled, a generalized profile of prescriptions may be constructed. In generalizing broadly, of course, important variations will be minimized and the generalized profiles may not be representative of a particular author or practitioner.

Emphasis has been placed on using test subjects to provide data for revision of prototype instruction. Test subjects are utilized after internal/expert review procedures have been conducted. Most recommend that both the developer and test subjects play active roles in the testing sessions. Test subjects should be representative of the target population. Both one-to-one and small group sessions should be conducted; test subjects should be debriefed. Both quantitative and qualitative (including attitudinal) data should be gathered. The importance of the analysis of posttest data is stressed. Cycles of testing and revision should continue until the instructional objectives have been met according to the designer's standards.

Research Questions

The prescriptions for developmental testing show variability among authors. Such a lack of reliability can be reduced only when sufficient research has accumulated which empirically demonstrates that a particular procedure is effective. The final criterion for effectiveness is, of course, how well and how efficiently the instruction functions.

Some of the procedures listed above raise some important questions. For the selection of a representative sample of test subjects, it remains unclear as to which characteristics/traits optimize test subjects. For example, does ability in this context refer to job performance, prior knowledge of the material to be learned, or is it the ability to learn? Clearly more research needs to be conducted on the impact of these and other subject characteristics on test data. Paralleling this issue, it remains

unclear as to how to sample from the target population in order to insure that ideal subjects are selected. More research is needed on test subject identification procedures.

Even areas of agreement among authors deserve empirical validation. It is almost universally recommended that attitudinal data be gathered but the most effective means of gathering such data, how it should be treated and what impact it should have on revision — all remain unexplored. Authors generally agree that developmental testing should be carried out on several subjects of "differing abilities." But the practitioner has little guidance from the authors in defining what is meant, for example, by "high, medium and low ability subjects". And, even worse, there is little research to support the assumption that different ability subjects (however defined) contribute differentially to the testing data. The seminal work of Wager (1983) deserves expansion.

In practice, formative evaluation may be carried out on instructional materials in other than the medium of text; see, for example, the extensive pre-testing of segments of the Sesame Street programs (Yelon, 1974) and the formative evaluation of instruction for the Open University (Brown, 1978; Henderson and Nathenson, 1976). Yet almost all of the research on developmental testing has been carried out using short text modules. In order to generalize those guidelines for testing which prove to be supported by empirical research, further studies should be carried out in various media.

Although revision heuristics and algorithms exist in the literature, there has been little research to support their validity and reliability. More research on the effectiveness of processing student test subject generated data (both qualitative and quantitative) through the procedures of well-defined revision strategies would be useful. Valid and reliable revision strategies have the potential to reduce much of the subjectivity that presently characterizes the revision process.

The final phase of the research programme being reported on in this dissertation addresses the need for research on the process of collecting feedback data from test subjects. The research questions raised above would require a number of research programmes in order to address all of the issues surrounding developmental testing and revision. Chapter V will report on an experiment designed to describe the comparative differences/similarities of one-on-one and small group testing and active versus passive roles for testers and test subjects. These differences and similarities will be described

in terms of the amounts and general kinds of data that are generated by using a number of endorsed procedures for designing and conducting developmental testing sessions.

Chapter V

An Experimental Manipulation of Developmental Testing Variables

Rationale for the Experiment

A basic rationale for doing formative evaluation, i.e., internal review, one-to-one, small group, and then field testing, appears to be related to the nature of design models. Models are presented as design heuristics. Applying a heuristic does not guarantee that the design goals/objectives will be attained. For this reason, it is prudent to review and revise an instructional design effort in order to monitor and improve the resulting instructional product, i.e., to formatively evaluate.

This chapter will describe the methodology of a study examining the differences and similarities of data obtained using the one-to-one and/or small group procedure for designing and conducting formative evaluations. Within the research literature related to developmental testing, the emphasis has been on comparing the effectiveness of instruction revised based on either one-to-one sessions, small group sessions, or on both one-to-one and small group sessions. This research strategy appears to address only the second of two critical research questions, related to formative evaluation, as formulated by Baker and Alkin (1973): "1. What are the major types of data to obtain and the major sources of such data? and 2. What is the impact of evaluation data upon the modification of programs?" (p. 393).

The effectiveness of revised versions of instructional material under varying experimental conditions has been a frequent dependent variable within the developmental testing of instructional materials research literature because the reason, in practice, for conducting testing sessions is to obtain data for purposes of revision. Although research, to this point, has tended to emphasize the relative effectiveness of revisions, the kinds of data produced from different testing procedures should not be ignored. It is quite feasible to suppose that the data produced under one-to-one and small group

conditions will differ in amount and/or kind, otherwise why bother conducting product tryouts using both strategies? Burt and Geis (1986) suggest that, within one-to-one/small group sessions, evaluators utilize a number of different sources for collecting feedback data: student test subjects articulating their problems while trying to learn from the material, posttest errors, affective reactions to a questionnaire, and debriefing results after the test session all can provide data. Differing roles that the participants could play during a testing session (e.g., an actively intervening and probing evaluator versus a more passive observer) might affect the data from these sources. However, no research until now has systematically described the data from these different sources and conditions. A better understanding of these data utilizing various formative evaluation methodologies should ultimately facilitate an understanding of the revision process — but this research does not look at the process of revision nor the effectiveness of revision.

The reviewed literature has not reached consensus on the relative effectiveness of one-to-one and/or small group testing — although both produce more effective instruction than the original. This is discussed in Chapters III and IV and summarized in Burt and Geis (1986).

The intent of this research is to describe the amounts and kinds of information (feedback) developmental testing provides, as a process utilizing a number of data collection methodologies.

Description of the Instructional Materials

Criteria Used to Select the Stimulus Materials Employed in the Experiment

The learning materials used in all experimental developmental testing sessions consisted of the first two units/chapters of the print module "Microbiology", one book in the series: Instructional Materials for the Dental Health Professions (Dental Auxiliary Education (D. A. E.) Project, 1982). These materials were available in rough (first) draft form and were selected because they:

1. were in draft form
2. were less than "perfect" (i.e., need revision)

3. were in replicable rather than ephemeral form (e.g., print rather than lecture)
4. required about an hour of learner time
5. concerned a subject matter that was unfamiliar to many people in the target population
6. did not require prerequisite knowledge that the target population would not have
7. had been developed by instructional designers applying the guidelines of systematic instructional development
8. provided criterion-referenced tests developed by the authors

The two Microbiology units met the above criteria and for this reason were selected. Unit One can be found in Appendix D and Unit Two in Appendix E.

A History of the Development of the Materials

The Microbiology module is only one component of a series of self-instructional modules that comprise a major part of a complete dental auxiliary curriculum designed to train professional dental assistants and dental hygienists. As described in the Faculty Guide for the complete set of curriculum materials, the Microbiology module is considered to be a sub-module within the curriculum because it covers clinical, basic science, or dental science information that is needed throughout the whole training program. In this respect, sub-modules cover information and/or skills that apply and generalize to more than one topic within the curriculum. A sub-module is considered to be a sub-topic of a larger module and in the case of the Microbiology module, it is a sub-topic of the broader topic: "Sterilization and Disinfection". The Microbiology module can be considered to be either a prerequisite to the topic of Sterilization and Disinfection or as an adjunct in which case a learner would refer to the Microbiology module while going through the Sterilization and Disinfection module.

The design of the D. A. E. project and resultant curriculum was stimulated by a grant from the W. K. Kellogg Foundation to the Division of Research in Medical Education of the University of Washington. The overall goal of the project was to: "develop and validate competency-based, self-instructional materials for students of the dental health professions and to implement a model of curriculum for which the materials were designed." (Microbiology: Instructional Materials for the Dental Health

Professions - D. A. E. Project, preface p. xv) The materials were developed using teams of content experts, instructional designers, and evaluators.

A modification of Dr. Robert Horn's (1976) Information Mapping system served as the basis for formatting the modules according to the subject matter and characteristics of the target population. Using this system results in the presentation of the information in blocks which are separated by lines in order to facilitate the efficient location of information. Beside each block of information, marginal labels are used to identify the subject and type of information. This is intended to aid the reader in scanning the module for the purposes of locating or reviewing important sections of information. Illustrations are used to aid in clarifying the printed material. Introductions are used at the beginning of each unit to preview the reason for learning the material. The objectives for each unit/chapter of any given module are presented to the learners in a detailed and behaviorally concrete manner following the introduction of each chapter. The grouping and sequencing of units within the modules are determined by the type of information being conveyed. Procedural information, information about equipment and materials, scientific and conceptual information are considered to be different kinds of information and therefore suggest different sequencing. The two units used in the experiment would be classified as scientific and conceptual information. With this kind of information, a separate section or block is used to introduce new terms, topics, or conceptual information in general. Whenever there are categorical schemes to be learned, learners are presented separate sections in a list format that explains relationships, diagrams, and outlines. For each chapter within a module, a criterion referenced self-test is included as an aid in helping the learner to identify particular strengths and weaknesses. A summary of the information presented is included just before each self-test in order to help the learner prepare for the self-test. Self-tests are designed to help learners prepare for a final posttest on the whole module. These posttests are formally administered and supervised by instructors. The learners must achieve at least 90% on the posttest and are given three chances to do so if they have failed to reach this criterion on earlier occasions.

A section in the Faculty Guide for using the materials summarizes the philosophy behind the use of self-instructional materials within the Dental Auxiliary Education (D. A. E.) Project:

The modules are designed to be used in a way that:

1. allows the student to acquire information about skills and basic and dental science independently of the instructor.
2. allows students to learn subjects and skills at their own rate without many of the constraints of traditional programs.
3. allows teachers to do what teachers do best: counsel, evaluate, diagnose, and prescribe for student learning. This frees them from the drudgery that is the mere telling of facts and procedures that are best communicated in other ways. (p. 10)

Differences Between Unit One and Unit Two of the Materials

One of the most important differences between Units one and two is that Unit one is prerequisite to Unit two. Unit one also serves to introduce the whole module of which the two units are only the first two sections.

Unit one presents a general description of four different kinds of microorganisms and various techniques used to study these microorganisms in a laboratory setting. The first unit presents a classification system for categorizing microorganisms into four different groups. One of these groups, bacteria, is broken down further categorically by shape. Three different shapes of bacteria are presented as examples of different kinds of bacteria. Drawings of these shapes are provided for the student. The general laboratory techniques used to study microorganisms are briefly discussed. The unit contains five stated objectives and presents five pages of self-instructional text. A self-test at the end of the unit contains five criterion items. Two of these items utilize a multiple-choice format; two items require fill-in-the-blank responses; and one question requires a student to recall a list of terms. When the five objectives of the first unit were classified according to the Bloom et al. (1956) taxonomy of educational objectives, one objective was classified at the knowledge level; three objectives at the comprehension level; and one objective at the analysis level of the cognitive domain of learning. When test items were similarly analyzed, all five were found to require responses at the knowledge level. Unit one presented 14 separate "blocks" of information according to Horn's (1976) information mapping system.

The second unit presents a general description of the environmental conditions necessary for the growth of microorganisms. Microorganisms normally present and not normally present in the human body are discussed. Two different types of infection are presented along with the factors necessary for an infection to develop. Three barriers that the human body has in order to defend itself from infection are discussed. The unit focusses on discussing two basic kinds of immunity systems that the body might use as a barrier to infection. The natural immunity system is discussed and two different kinds of natural immunity are introduced. The acquired immunity system is discussed under two major sub-classifications. These sub-classifications are then further divided into two more sub-groupings. A categorical scheme (a figure/diagram) is presented as an aid for the student in comprehending the immunity system and its sub-systems. The unit provides instruction for six stated objectives (although seven are listed) and presents eight pages of self-instructional text. The original developers of the instruction had decided to provide instruction for the seventh objective in the third unit of the module. This simply highlighted the rough draft state of the materials used in the experiment. The self-test at the end of the second unit contained 14 criterion items. Only 12 of the 14 items were used in this study. Two items which were not keyed to the instruction were discarded by the author of this study. Of the 12 items which were used, two of these items utilized a multiple-choice format; seven items required fill-in-the-blank responses; and three questions required a student to recall lists of terms. When the six objectives of the second unit were classified according to Bloom et al. (1956) taxonomy of educational objectives, three objectives were classified at the knowledge level and three objectives at the comprehension level of the cognitive domain of learning. When test items were similarly analyzed, seven items were found to require responses at the knowledge level; three items at the comprehension level, and two items at the synthesis level.

Unit two contains 21 separate blocks of information. These blocks also tend to contain more information within them than the 14 blocks of information in unit one. Unit two is much longer than the first unit. More scientific terminology is presented and more categorical schemes are to be learned than in the first unit.

Description of the Population Utilized and Sampling Procedures

Of critical concern to this study was the matching of the materials with an appropriate target population. The materials had been originally designed for dental hygiene students but it was discovered that there were too few dental hygiene programmes (offering instruction in the English language) from which to select a suitable number of test subjects. A curriculum development specialist at McGill University, familiar with the college-level nursing curriculum, reviewed the materials and suggested that the materials would suit first year nursing students. Of the three junior colleges offering nursing programmes in English on the Island of Montreal, two were randomly selected. Both the chairpersons and academic teaching staff of these two programmes reviewed the instructional materials and judged them to be suitable for the first year students. None of these students had been exposed to Microbiology within their programmes as this was a course taken in the second year of study. As a precaution, the subjects who participated in the study were asked to provide background information about their prior knowledge of the material used in the experiment. It was reasoned that some might have been exposed to the material used in the experiment through participation in other programmes, reading books, etc. Table 9 illustrates how some of the sample acquired prior knowledge of the material used in the experiment.

The two junior colleges chosen to participate in the experiment were located in different neighborhoods on the Island of Montreal. One college was located in the inner-city and was reported to draw its student population from largely lower socio-economic groups. The other college was located in a much more affluent neighborhood and reportedly draws its student population from higher socio-economic groups. These statements about the student population's SES are based on anecdotal comments — no data were obtained on SES and it was not considered as a variable in the study. It was reasoned that since the two colleges offered very similar programmes (due to government and professional nursing associations' setting of the nursing curriculum for certification purposes) and required the same admission requirements that both colleges would provide members of the target population for the instructional materials. Administrators of the two junior colleges were promised that in no way would

comparisons be made between the two colleges. Anonymity was guaranteed in order to secure the cooperation of the administrations.

Two campuses were canvassed for volunteers. Test subjects were informed of the nature of the study and were told that they would each receive a \$50.00 stipend for their time. Appendix F contains the consent form which volunteers were required to sign in order to participate in the experiment. It briefly describes what was being requested of the volunteer. From each college thirty-six test subjects were randomly selected from the volunteer pool. A total of seventy-two subjects were utilized in the study: three of the seventy-two subjects were male and sixty-nine were female. Table 6, table 7, table 8, and table 9 present demographic information that was collected from the subjects during the experiment. Specific questions soliciting this background information were embedded into the "package" of stimulus materials.

In addition to collecting demographic data, the calendars for the two nursing programmes were reviewed. A review of the descriptions of the two programmes revealed very similar programmes along a number of important considerations: admission standards, specific academic background requirements for students, coursework to be completed, and standards to be met in order to successfully complete the programme.

Experimental Design and Research Questions

The following descriptive research questions are addressed in this study:

1. Do experimental conditions affect the amount of data obtained from the different sources of feedback? and Do experimental conditions affect the amount of data obtained from the combined sources of feedback?

Sources: a) comments made during instruction (including any comments made while responding to individual self-test items)
 b) debriefed comments
 c) debriefed self-test errors
 d) debriefed attitudinal survey results

2. Does the test subject characteristic of prior knowledge of the material affect the overall amount of comments being made during instruction?

Table 6.

Distribution of Samples' Mother Tongue

<u>Tongue</u>	<u>Frequency</u>	<u>Percent</u>
English	47	65.3
French	8	11.1
English and French	1	1.4
East Indian	1	1.4
Tagalog (Filipino)	3	4.2
Greek	4	5.6
Polish	3	4.2
Spanish	2	2.8
Italian	1	1.4
Hungarian	1	1.4
Portuguese	1	1.4
Total	72	100.0

Table 7.

Distribution of Samples' Age

<u>Years of Age</u>	<u>Frequency</u>	<u>Percent</u>
17	10	13.9
18	15	20.8
19	7	9.7
20	7	9.7
21	5	6.9
22	6	8.3
23	4	5.6
24	4	5.6
25	1	1.4
26	2	2.8
27	1	1.4
28	4	5.6
29	1	1.4
31	1	1.4
32	1	1.4
34	1	1.4
37	1	1.4
40	1	1.4
Total	72	100.0

(Mean Age = 21.625; SD = 4.935)

Table 8.

Distribution of Samples' Educational Background

<u>Years of Schooling</u>	<u>Frequency</u>	<u>Percent</u>
11	21	29.2
12	14	19.4
13	11	15.3
14	11	15.3
15	8	11.1
16	2	2.8
17	4	5.6
18	1	1.4
Total	72	100.0

Table 9.

Distribution of How Some of the Sample Acquired Prior Knowledge

<u>Prior Knowledge Through:</u>	<u>Frequency</u>	<u>Percent</u>
No Prior Knowledge	34	47.2
A Course in Microbiology	9	12.5
Reading Books	14	19.4
Work or Work Related Activities	4	5.6
A Nursing Course	2	2.8
Reading Books & Work Related Activities	5	6.9
Work Related Activities & a Lab on Asepsis	1	1.4
Grade 9 Biology	1	1.4
A Hairdressing Course	1	1.4
Missing Data	1	1.4
Total	72	100.0

(Prior Knowledge: N= 37)

3. To what extent do data from the various sources of feedback provide redundant or unique information? This is a general question addressed by more specific questions:

Specifically: a) After instruction, as a debriefing strategy, to what extent does the procedure of having test subjects go over their comments about the materials again (in retrospect) produce unique or redundant feedback data?

b) To what extent does performance on individual self-test items reflect comments (perceived problems) made during instruction?

c) What are test subjects' perceptions of self-test error causation?

d) After instruction, as a debriefing strategy, to what extent does the procedure of having test subjects review their self-test errors produce unique or redundant feedback data?

e) After instruction, as a debriefing strategy, to what extent does the procedure of having test subjects review their unfavorable attitudinal survey responses produce unique or redundant feedback data?

4. What brings a comment made during instruction to the debriefing agenda, e.g., do subjects just locate and not explain the problem or can they explain and/or remediate the problem?
5. What general and specific kinds of problem explications and revisions/revision prescriptions result from developmental testing and debriefing?
6. Do experimental conditions affect the amounts of general kinds of problem explications and revisions/revision prescriptions made during instruction?
7. Do experimental conditions in conjunction with the test subject characteristic of prior knowledge of the material affect the amounts of general kinds of problem explications and revisions/revision prescriptions made during instruction?

Figure 7 presents the experimental design utilized to collect the data. Subjects were randomly assigned to either one-to-one or small group conditions (the independent variable of number of participants) and to either active or passive conditions (the independent variable of role of participants). The independent variable of number of participants is similar to the size of class variable in the classroom processes literature.

This design permitted the analysis of variability of the frequency data through a 2×2 ANOVA. Frequencies were tabulated for the number of responses (amount of data), by test subject, for each one of the four sources (dependent variables) of feedback data: 1) the number of comments made during

Experimental Conditions		Sources of Developmental Testing Feedback Data					
		During Instruction			Debriefing		
Active Roles	One-to-One N=18	comment on material & self-test	take self-test	respond to survey	discuss comments	discuss self-test errors	discuss survey responses
	3 Small Groups N=18	comment on material & self-test	take self-test	respond to survey	discuss comments	discuss self-test errors	discuss survey responses
Passive Roles	One-to-One N=18	comment on material & self-test	take self-test	respond to survey	discuss comments	discuss self-test errors	discuss survey responses
	3 Small Groups N=18	comment on material & self-test	take self-test	respond to survey	discuss comments	discuss self-test errors	discuss survey responses

Figure 7. The procedures used within the experimental design. The design was applied to Unit one and replicated on the second day with Unit two.

instruction including any comments related to the self-test items, 2) the number of comments about the instruction made during debriefing, 3) the number of errors made on the self-test which were debriefed, and 4) the number of unfavorable responses to the attitudinal survey which were debriefed. For each of the independent variables it was assumed that test subjects' comments made during instruction, comments made during debriefing, self-test errors, and attitudinal responses would be normally distributed, e.g., a few subjects would make many comments, most subjects would make some comments and a few subjects would make no comments etc.

This design provided the test subjects a number of different opportunities to make comments about the materials and self-tests: 1) all subjects were requested to make comments on the materials and self-tests during instruction; 2) after instruction, during debriefing, test subjects were requested to review these comments in order to finalize the results based on this source of obtaining feedback data; 3) then subjects were informed of the results of the self-tests. These results were discussed in order to obtain more comments about the materials and self-tests from the test subjects; 4) finally test subjects were informed of the results of the attitudinal survey. These results were discussed in order to obtain more comments about the materials and self-tests from the test subjects. A content analysis, based on all of these comments made by all of the test subjects, was conducted in order to describe both the specific and general kinds of comments (problem explications and revisions/revision prescriptions) that test subjects made with regards to the self-instructional materials used in the experiment.

Figure 8 presents a 2 x 2 MANOVA data analysis design with six dependent variables. The independent variables for this design were the same as the 2 x 2 ANOVA design. The dependent variables represent tabulated frequencies of the general kinds of responses (comments) which test subjects made during instruction. Figure 9 presents a 2 x 2 x 2 MANOVA data analysis design with the same six dependent variables. This design added the additional factor of prior knowledge/no prior knowledge to the analysis. Both of the MANOVA data analysis designs were devised in order to look at the effects of the independent variables on the amounts of the general kinds of comments that test subjects made during instruction.

DEPENDENT VARIABLES (Categories of Responses)

		Category #1	Category #2	Category #3	Category #4	Category #5	Category #6
INDEPENDENT VARIABLES	Active Roles	One-to-One (N=18)					
		3 Small Groups (N=18)					
	Passive Roles	One-to-One (N=18)					
		3 Small Groups (N=18)					

Categories of Responses:

1. Comprehension/Recall kinds of problem explications and revisions/revision prescriptions
2. Motivation/Attention kinds of problem explications and revisions/revision prescriptions
3. Physical Feature kinds of problem explications and revisions/revision prescriptions
4. Multiple Responses overlapping two or more of the above three kinds of problem explications and revisions/revision prescriptions
5. Non-Classifiable Responses for kinds of problem explications and revisions/revision prescriptions
6. Responses indicating that the test subject was unable to explicate a problem and/or suggest a revision after a problem had been located

Figure 8. 2 x 2 MANOVA data analysis design #1 with six dependent variables. Unit #1, Unit #2, and totals for the combined units were analyzed separately for both problem explication and revision responses.

DEPENDENT VARIABLES
(Categories of Responses)

			*1	*2	*3	*4	*5	*6
INDEPENDENT VARIABLES	Active Roles	Prior Knowledge	One-to-One (N=9)					
			3 Small Groups (N=11)					
		No Prior Knowledge	One-to-One (N=9)					
			3 Small Groups (N=6)					
	Passive Roles	Prior Knowledge	One-to-One (N=9)					
			3 Small Groups (N=5)					
		No Prior Knowledge	One-to-One (N=9)					
			3 Small Groups (N=13)					

Categories of Responses:

1. Comprehension/Recall kinds of problem explications and revisions/revision prescriptions
2. Motivation/Attention kinds of problem explications and revisions/revision prescriptions
3. Physical Feature kinds of problem explications and revisions/revision prescriptions
4. Multiple Responses overlapping two or more of the above three kinds of problem explications and revisions/revision prescriptions
5. Non-Classifiable Responses for kinds of problem explications and revisions/revision prescriptions
6. Responses indicating that the test subject was unable to explicate a problem and/or suggest a revision after a problem had been located

Figure 9. 2 x 2 x 2 MANOVA data analysis design #2 with six dependent variables. Unit #1, Unit #2, and totals for the combined units were analyzed separately for both problem explication and revision responses.

This study will focus on the results of the data analyses and present a categorical framework based on a content analysis of all of the comments which test subjects made during instruction and debriefing. The content analysis was used to generate six general categories (the dependent variables used in both of the MANOVA statistical designs) of the kinds of comments that student test subjects made. Although the results of the content analysis are presented in chapter 6, the six general kinds or categories of comments/responses (the dependent variables) are listed below in order to both provide a brief overview and to insure that Figures 8 and 9 are clear:

1. Comprehension/Recall kinds of problem explications and revisions/revision prescriptions
2. Motivation/Attention kinds of problem explications and revisions/revision prescriptions
3. Physical Feature kinds of problem explications and revisions/revision prescriptions
4. Multiple Responses overlapping two or more of the above three kinds of problem explications and revisions/revision prescriptions
5. Non-Classifiable Responses for kinds of problem explications and revisions/revision prescriptions
6. Responses indicating that the test subject was unable to explicate a problem and/or suggest a revision after a problem had been located

Two types of comments emerged from the data collection after problems had been physically located. The first type related to the general kinds of problems which student test subjects perceived during instruction, e.g., comprehension problems, motivation problems, etc., while the second type related to the general kinds of revisions/revision prescriptions that student test subjects generated during instruction, e.g., change the physical features of the material, add motivation, etc. For each of the dependent variables it was assumed that test subjects' responses would be normally distributed, e.g., a few subjects would have many comprehension/recall problems, most subjects would have some of these problems and a few subjects would have none, etc.

Study Procedures

From each group of thirty-six test subjects, subjects were randomly assigned to one of four experimental conditions: 1) one-to-one active; 2) one-to-one passive; 3) small group active; or 4) small group passive. Eighteen one-to-one active sessions ($N=18$), eighteen one-to-one passive sessions ($N=18$), three small group active sessions with six subjects per small group ($N=18$), and three small group passive sessions with six subjects per small group ($N=18$) were conducted by the same experimenter.

All conditions. For both the one-to-one and small group experimental conditions, subjects were instructed that the materials were in first draft form and that the purpose of developmental testing was not to test how well subjects could learn, but to learn how well the materials taught. All subjects were told that their feedback on the effectiveness of the material was important. Test subjects were asked to locate the problems they were encountering (any kind of problem), to explain the nature of the perceived difficulty, and to suggest, if possible, a revision addressing the problem so that another student would not encounter the same problem. For each condition subjects were given specific instructions for purposes of task orientation. These instructions defined the role that the subject(s) was to play during developmental testing. Subjects were requested to read the instructions as the experimenter read them out loud. All subjects were asked to leaf through the instructional material, while the experimenter instructed them to attend to the self-test and the attitudinal survey. They were then told that there would be a short break after which they would be debriefed (interviewed) based on their responses to the material. A tape recorder was used to record the interactions between the test subject(s) and the experimenter. After instructions were discussed, subjects were asked if they had any questions as to what was being expected of them. Appendix G contains all of the instructions presented to subjects in all four of the experimental conditions described below.

One-to-one active. During active one-to-one developmental testing, each subject was instructed to go through the materials as best they could and if either the subject or the experimenter perceived that there was a problem with the materials it would be discussed during the session. Cues or signals

indicating that a problem was perceived included either verbal behavior: the test subject making some kind of a spontaneous comment or a subject's response to the experimenter's page-by-page probe: "Were there any problems on that page?" ; or non-verbal test subject behavior signalling to the experimenter that there might be a problem (i.e., rapid page-turning, fidgeting, or when the subject started highlighting/writing on the material). Subjects were given a yellow highlighting marker and instructed to use it in order to identify/locate their perceived problems by underlining the problematic portion(s) of the textual display. If a test subject indicated that there was a problem or if the experimenter perceived there to be a problem, the experimenter responded by saying to the individual test subject either:

1. "What is the problem?" or
2. "Subject's name, I've noticed that you have been (e.g., fidgeting). Are you having a problem?"

Under either circumstance, the experimenter probed the test subject until there was a definite indication of a problem. If a test subject persisted in asking content questions of the experimenter, he would respond by saying: "Where were you in the material when you began to feel that you had a problem?" The experimenter did not provide instructional assistance, only probing whenever difficulties were observed. In all circumstances the subject would locate the problem area resulting in either the experimenter or the subject underlining the portion(s) of the text with the highlighter. Then the experimenter probed further by saying: "What is the problem?" Thus, before going on, the problem was explicated by the test subject resulting in either the subject or the experimenter writing on the text the nature of the perceived problem. Upon explication of the problem, the test subject was asked to suggest a way of avoiding that problem for the next student who would encounter the instructional material. The experimenter, in order to operationalize a revision "set", responded by saying: "What could be changed within the material so that somebody else wouldn't have that problem?" If the subject was unable to generate a revision then he/she was asked to move on within the instructional material. These experimental "instructions" were written on a cover sheet for the materials and were verbally explained at the onset of the experiment.

One-to-one passive. During passive developmental testing, subjects were instructed to go through the materials as best they could, use the highlighter to locate any problem area(s), use a pen provided to explain the problem by writing on the material itself, and finally, if possible, to make/suggest a revision. After explaining the task for the subject, the experimenter left the room in order to let the subject work alone.

Small group active. During small group active developmental testing, subjects were given identical instructions as described for one-to-one active. The only differences between the conditions were that subjects were requested not to talk with their "neighbors" during the session as there were now six test subjects sitting around a large table with the experimenter. Subjects were requested to raise their hand if they wished to express that they perceived a problem. After explaining the problem to the experimenter, the subject's response was written on the instructional material.

Small group passive. For small group passive developmental testing, subjects were given identical instructions as described for one-to-one passive. Similarly to small group active, subjects were requested not to talk with their "neighbors" during the session. The experimenter did stay in the room with these subjects but did not interact or probe subjects as difficulties arose. These subjects were individually on their own locating problems with the highlighter, explaining the problems on the materials in writing, and if possible, making/suggesting revisions in writing.

Instrumentation and Data Sources

The experimental design utilized in this study called for the testing of an individual subject or a group of subjects over a consecutive two-day period. On day one, the first unit of the material was developmentally tested. On the second day, the second unit of the material was tested.

In addition to test subjects identifying/locating perceived problems, explicating why they perceived something to be problematic and suggesting ways in which that problem could be remedied, they were also requested to complete a criterion-referenced self-test (they were permitted to criticize the self-test as well), and complete an attitudinal survey. Both the criterion referenced self-tests and the attitudinal survey can be found in the experimental test subjects' "packages" of instructional

materials located within Appendices D and E. The self-tests were designed based on the objectives expressed within the material and the tests sampled the attainment of specific objectives for which subjects had received instruction. Most items on the tests sampled a subject's recall of the content. In general, higher levels of comprehension/ performance were not tested. The test for Unit one contained five items using both multiple-choice and fill-in-the-blank formats. The test for Unit two contained 12 items (reflecting a longer unit than the first) using multiple-choice, fill-in-the-blank, and short-answer formats.

A longer test was also contained within the module. This test was designed to function as a criterion-referenced self-test and would be taken by a student after completion of the complete module. As only two units of instruction were experimentally developmentally tested in this study, it would have been inappropriate to have used the longer self-test.

An adapted version of Abedor's (1972) attitudinal survey was utilized in this study. The original survey had been designed for use within a multi-media instructional delivery system, (i.e., a self-instructional print workbook in combination with audio-visual media). The adapted version reflected the use of a self-instructional print module as the medium of instructional delivery. The same adapted version was used for the two instructional units. The adapted version contained 27 items of which 23 were on a five point Likert-type scale ranging from strongly agree to strongly disagree. Four items were of open-ended format. Overall, the items were general in nature, requiring a subject to globally evaluate the instructional material. Statements on which students would base their ratings (the 23 Likert-type scaled items), reflected the major decisions that are made during an instructional design effort using a systematic design model.

Criteria Used to Construct the Debriefing Agenda

After having completed a unit of the materials, including the self-test and the attitudinal survey, the subject(s) were given a 15-20 minute break. During this time the experimenter compiled an agenda for purposes of debriefing test subjects based on the data which were collected during

instruction, i.e., the written comments the subject made on the materials and self-tests, the performance data tabulated from the self-test, and the tabulated attitudinal survey response data.

Criteria for an item appearing on the debriefing agenda varied depending on experimental condition. For one-to-one conditions, all problems identified (test subject comments) while going through the material including comments made while writing the self-test, all errors made on the self-test, any response to a Likert-type scaled item that was two points away from the ideal response and unfavorable open-ended survey responses were put on the agenda for discussion during debriefing. All of the outcomes of the discussions were transcribed from the tape recordings.

For small group conditions, if two or more test subjects perceived a problem within the same section of the instruction (including comments made while completing the self-test) then it was placed on the agenda for debriefing. If two or more test subjects made an error on the same self-test item, the item and the section of the material which provided instruction for that particular item became an agenda item for debriefing. If the groups' mean response to an attitudinal survey item was two or more points away from the ideal response, that item was discussed during debriefing. A similar procedure was applied to the open-ended questions — if two or more subjects responded unfavorably, it became a debriefing agenda item. After small group developmental testing, an assistant aided the experimenter in tabulating the data collected during the session for purposes of generating the debriefing agenda. Forms were created in order to facilitate the rapid tabulation of the small group developmental testing data in matrix format. These forms were designed based on revision guidelines recommended by Dick and Carey (1985) and can be found in Appendix H.

After the break, the experimenter debriefed the test subject(s) using a tape recorder to record the session. Debriefings began with the experimenter providing an overview of the purpose of the debriefing. Subjects were also told how the agenda was created. In all conditions the problems relating to the text and the test that had been identified, explicated, and/or revised during instruction were reviewed first in order provide test subjects the opportunity to further comment on the nature of their problems and revisions. Subjects were permitted to change their minds, identify and resolve problems that had not been detected during instruction, or verify that what had been a problem during instruction

was still a problem after instruction. A similar procedure was then applied to self-test item errors. For this component of the agenda, subjects were asked to consider if the problem (an error on the self-test) was related to: the wording of the item, the section in the unit providing instruction for that item, both the item and the instruction, or that the error was the result of negligence on the part of the test subject(s) in which case the subject(s) would agree that there was no problem with the materials. After agreement was reached that there was a problem, subject(s) were requested to explicate and resolve the problem. The final agenda component addressed unfavorable responses to both the Likert-type scaled items and the open-ended items on the attitudinal survey. Subjects were permitted to change their minds, identify and resolve problems that had not been detected during instruction, or verify that what had been a problem during instruction was still a problem after instruction. This debriefing procedure allowed the subjects to elaborate on why they had indicated dissatisfaction on any given item as well as suggest resolutions addressing the problem.

Different Debriefing Strategies Utilized Between One-to-One and Small Group

A major difference between the one-to-one and small group conditions is the degree of interaction afforded through debriefing. In one-to-one, all problems identified from all of the sources of developmental testing data were discussed between the experimenter and the individual subjects during debriefing. All of the outcomes of the discussions were transcribed from the tape recordings. As previously discussed, this was not the case during small group due to the criteria imposed in order to construct the debriefing agenda. When an item appeared on the small group agenda, it was discussed by all of the members in the group ($N=6$). Individual group members were allowed to openly discuss the identified, explicated, and resolved problems respectively between themselves. The experimenter functioned as a moderator for the discussion. It should be recalled that during instruction at least two group members must have indicated a problem within the same section of the materials; at least two members had to have made an error on a self-test item; and that the groups' mean attitudinal survey response was used to construct the debriefing agenda. Due to the imposed criteria for setting a small

group agenda, there were times when individual group members were requested to comment on agenda items with which they personally had had no problems. This applied to the three components of the agenda: 1) comments and suggestions directed towards the text and the test items during instruction; 2) self-test errors; and 3) unfavorable reactions to the attitudinal survey. During the discussions held during the debriefings, the experimenter would try to get the group to reach a certain amount of agreement on where the problem was located, what the problem was, and what to do to resolve the problem. Individual subjects within the group were permitted to indicate that they still had no problem with any given agenda item. In all cases, small groups were able to reach agreement, as previously defined, on all of the items which appeared on all of the agendas based on all of the feedback sources.

During the debriefing of a small group agenda item, the experimenter would randomly choose a subject to begin the discussion. This insured that all subjects would have opportunities to start the discussion. The experimenter would bring an item up for discussion by stating that some members of the group had indicated that there was a problem — the experimenter did not identify those individuals within the group. He reminded the group members that there might be times when only one individual had indicated a problem and therefore it would not be discussed. This was done in order to insure that test subjects understood that their comments and suggestions were still valuable and welcomed. The experimenter directed the groups' attention to the relevant section within the material, or self-test item, or attitudinal survey item. If at least two of the group members maintained that there was a problem then the group was asked to explicate that problem. Then the group was directed towards resolving the problem.

Debriefing can be conceptualized as a retrospective data collection procedure used to finalize the outcomes of developmental testing for a given individual test subject or group of test subjects. After debriefing the first unit, the test subject(s) was asked to reconvene for the next day to developmentally test the second unit of the materials. The notes taken during debriefing were revised and elaborated through review of tape recordings of the debriefing sessions.

In review, referring to Figure 7, for each of the experimental conditions for both Unit one and Unit two, data were gathered from the:

1. comments made during instruction (plus any comments made while responding to individual items)
2. self-test responses
3. attitudinal survey responses, and
4. debriefing based on:
 - a) comments made during instruction (including any comments made while responding to individual self-test items)
 - b) self-test item error data
 - c) unfavourable attitudinal survey response data

Relations were explored between all of the following dependent variables:

1. comments made during instruction and comments about the instruction made during debriefing
2. comments made during instruction and performance on the self-test
3. self-test results and test subjects' perceptions of self-test error causation
4. comments made while taking the self-test and comments made after knowledge of results of the self-test
5. comments based on knowledge of attitudinal survey results and any previous feedback

Across all sources of feedback, i.e., the comments during instruction, the comments after the instruction, debriefing the self-test results, and debriefing the attitudinal survey results, three different kinds of responses were elicited from test subjects:

1. problem identification/location
2. problem explication, and
3. revision/revision prescription

The manner by which the data (these three different kinds of responses) were collected and recorded varied between instruction and debriefing. During instruction, all test subjects were requested to

identify, explicate, and remediate perceived problems with the instruction and the self-test tests by actually writing on the instructional materials.

After instruction, during the debriefing, test subjects were requested to retrospectively identify, explicate, and remediate perceived problems with the instruction and the self-tests based on their comments made during instruction. Next on the agenda during the debriefing, test subjects were informed of the results of the self-test and they were requested to identify, explicate, and remediate perceived problems with the instruction and/or the self-test. Finally, test subjects were informed of the results of the attitudinal survey. These results had signalled, to the experimenter, that problems of a general nature had been identified (recalling that the attitudinal survey required a subject to globally evaluate the major components of the instruction). Test subjects were then requested to specifically identify, explicate, and remediate perceived problems with the instructional materials. This was done for all of the responses to the attitudinal survey indicating dissatisfaction. All data generated during debriefing were tape recorded for transcription purposes.

Results of Pilot Testing the Experimental Procedures

Each one of the four experimental conditions (one-to-one active, one-to-one passive, small group active and small group passive) was pilot tested in order to insure that subsequent experimental subjects assigned to these experimental conditions would not encounter difficulties due to unanticipated procedural problems. The pilot testing took one week to complete. Each of the participants in the piloting went through the developmental testing and debriefing of the two units of the prototype instructional materials over two days. One subject was used in the pilot testing of the one-to-one active experimental condition; one subject was used in the pilot testing of the one-to-one passive experimental condition; six subjects were used in the pilot testing of one small group active experimental condition; six subjects were used in the pilot testing of one small group passive experimental condition. The following changes were made to the experimental procedures based on the results of pilot testing:

1. After having pilot tested the passive small group, it was discovered by the test subjects that there was no accompanying instruction for the last two items of the self-test for the second unit. Subsequently to this group, all test subjects were instructed to delete these questions.
2. After having pilot tested the one-to-one passive test subject, it was decided that the experimenter would not stay in the room while the subject was going through the material. After the debriefing for the second unit of the material the test subject had indicated to the experimenter that she had felt uncomfortable with the experimenter being there. The experimenter experienced similar discomfort in this condition because there was nothing for him to do and felt there to be really no need in being there. Subsequently, all one-to-one passive experimental condition test subjects were left on their own to work through the prototype instructional materials. Subjects would be debriefed afterwards. Allowing the test subject to be alone in the room while going through the material as best they could simulated the most realistic conditions for learning from self-instructional material.
3. While pilot testing the first one-to-one active test subject, the experimenter felt that it would be desirable to ask the test subject, after completing each page of instructional text, if they had experienced any difficulty within the material. This would help maintain a "revision set" to keep the test subject focused on the problem identification, explication, and remediation tasks. This procedure was conducted for all subsequent one-to-one active test subjects

All of the data collected during pilot testing were used in the final data analysis.

Segmenting the Instructional Materials for Coding Purposes

The two units of the instructional materials were segmented into sections in order to facilitate the coding of test subjects' comments which took the form of: problem identifications, explications, and remediations. This permitted the keying of comments and suggestions to specific sections of the instructional materials. Sections were created to parallel the pre-existing divisions (blocks of information) within the materials. Blocks of information result from the utilization of Horn's (1976) Information Mapping system for instructional design. At times, it was decided that some of the blocks, for coding purposes, were too large. Within these larger blocks of information, several concepts, definitions, or lists of terms were being conveyed. As subjects who participated in the experiment tended to direct their comments and suggestions to specific concepts, definitions, etc., the larger blocks of information were broken down in order to facilitate the pinpointing of subjects' comments and suggestions within the materials. The individual self-test items were also assigned section numbers. The first unit of the instructional materials was divided into 29 sections while the second unit of the materials was divided into 50 sections.

Coding the Amount of Data Obtained from the Different Sources of Feedback

In order to address the research question: Do experimental conditions affect the amount of data from the different sources of data?, each source was conceptualized as a separate dependent variable. Each of the sources of data will be presented and discussed in relation to how the data was tabulated.

Source: Comments Made During Instruction

Each of the subjects who participated in the study was exposed to each of the 79 sections of instructional material. Test subjects were encouraged to freely comment on any aspect of the material which they perceived as being problematic. This procedure applied to each self-test item, as well, as

the items were an integral component of the materials. Self-test items were treated just like any other section of the materials and were assigned section numbers. For each subject, data were coded by individual section. For each of the 79 sections, individual subjects were coded as either having perceived a problem (coded as 1) or not having perceived a problem (coded as 0).

For all subjects, the total number of identified problems identified during instruction were tabulated, by individual subject, for both the first and second units of instructional material. In addition, the total number of identified problems across the two units was tabulated for each subject.

For the total number of identified problems (Unit one + Unit two outcomes), the prior knowledge variable was utilized to address the research question: Does the test subject characteristic of having indicated prior knowledge of the material affect the overall amount of comments made during instruction? Prior knowledge was not assessed by a pretest measure but was based on test subject self-report (either yes or no) after having completed the second unit. The following question was embedded just after the self-test for the second unit of instructional material. In the past, have you ever been exposed to subject matter content related to the topics of sterilization and disinfection in Microbiology? Thirty-four subjects indicated that they had prior knowledge, thirty-seven subjects indicated that they had no prior knowledge; there was missing data for one subject.

Source: Debriefed Comments

The second source of feedback data related to the problems which were identified (by test subjects commenting on the materials and self-tests) during instruction. Test subjects were asked during debriefings, in retrospect, to confirm, elaborate, or modify their comments made during instruction. Subjects were also permitted, during debriefings, to raise comments which had not been perceived during instruction.

It should be recalled that the criteria varied between one-to-one and small group conditions for determining the debriefing agendas based on comments made during instruction. All problems that were identified during one-to-one testing were debriefed. For a problem to be debriefed for small groups, at

least two members of a group had to have identified a problem within the same section of the material. The problem statements did not have to be similar for agenda setting purposes; if at least two group members had identified any kind of problem within the same section of material the problem section was discussed during debriefing.

For all subjects, the total number of comments which were debriefed based on the results of developmental testing were tabulated, by individual subject, for both the first and second units of instructional material. In addition, the total number of debriefed comments for the two units combined was tabulated for each subject.

Source: Debriefed Self-Test Item Errors

The third source of feedback data were the self-tests. Responses to individual items on the self-tests provided error data. Self-test item error data was used as an indicator of a possible problem within the materials and/or self-test item. All responses by test subjects to individual items were scored as either being correct or incorrect. Self-test error data were gathered during developmental testing and used to construct part of the debriefing agenda.

Response formats between the items on the self-tests varied. Due to this variability, items were scored differently at times. Table 10 presents the scoring key which was used to determine whether an item response was to be considered correct or incorrect.

Table 10.

Scoring Key by Individual Self-Test Item

<u>Unit #</u>	<u>Item Number</u>	<u>Item Scored as Incorrect if:</u>
1	1	at least one response incorrect
1	2	item response incorrect
1	3	item response incorrect
1	4	item response incorrect
1	5	item response incorrect
2	1	at least one response incorrect
2	2	item response incorrect

2	3	item response incorrect
2	4a	item response incorrect
2	4b	item response incorrect
2	5	item response incorrect
2	6	item response incorrect
2	7	at least one response incorrect
2	8	at least one response incorrect
2	9	item response incorrect
2	10	item response incorrect
2	11	item response incorrect
2	12	item response incorrect

It should be recalled that the way in which debriefing agendas were set differed between the one-to-one condition and the small group condition. There was a criterion imposed on the self-test item error data gathered from small groups which was not applied to the one-to-one self-test item error data. If two or more test subjects in a small group made an error on the same self-test item then that item and the section of the material which provided instruction for that particular item became an agenda item for debriefing.

For all subjects, the total number of debriefing agenda items which was based on the results of the self-tests was tabulated, by individual subject, for both the first and second units of instructional material. This was also done for the totals of Units one and two combined.

Source: Debriefed Attitudinal Survey Results

The final source of feedback was the attitudinal survey. Responses to individual items on the surveys provided an indication of the degree of satisfaction/dissatisfaction to various components of the instructional materials. Responses to individual attitudinal survey items indicating dissatisfaction were used as indicators of a possible problem within the materials. Attitudinal survey data were gathered during developmental testing and used to construct a part of the debriefing agenda.

It should be recalled that the way in which debriefing agendas were set differed between the one-to-one condition and the small group condition. Mean responses were computed for the Likert-type scaled items for small groups. If a group's mean response to a scaled item was two or more points away

from the ideal response, then the particular item was brought to the debriefing agenda. For open-ended attitudinal survey items, at least two members of the group had to have responded unfavorably in order for the items to be brought to the agenda. For one-to-one, an individual's response had to have been two or more points away from the ideal on a scaled item. Any unfavorable response to an open-ended item was brought to the debriefing agenda for one-to-one.

For all subjects, the total number of debriefing agenda items which was based on the results of the attitudinal survey were tabulated, by individual subject, for both the first and second units of instructional material. This was also done for the totals of Units one and two combined.

The Total Amount of Feedback Obtained

For each test subject, the total number of problems which was identified based on debriefing the comments, the self-test results, and the attitudinal survey results, i.e., the total number of debriefing outcomes, was tabulated by combining the total number from Units one and two. This would provide a general indicator of the total amount of feedback, collapsing across sources, obtained from a test subject.

Relationships Between the Various Sources of Feedback Data

In order to address the general research question: To what extent do the data from the various sources of feedback data provide redundant or unique information?, the relationships between the dependent variables (the sources of feedback data) were examined for descriptive purposes. The methods used to code/tabulate the data are described below for the specific research questions relating to the general question which guided this aspect of the research.

Relating Comments Made During Instruction to Comments About the Instruction Made During Debriefing

All of the comments (problem explications and revisions) generated during instruction were coded as either being redundant or not with the comments made during debriefing. For both of these sources of comments test subjects were requested to comment both on the material and the self-tests. As previously mentioned for small groups, there were criteria operating for putting the comments on the agenda. All subjects within a group were coded with the group's agreed upon response for both problem explications and revisions. The following research question was addressed: After developmental testing, as a debriefing strategy, to what extent does the procedure of having test subjects go over their comments about the materials again (in retrospect) produce unique or redundant feedback data?

Relating Self-Test Performance to Comments Made During Instruction

Each self-test item response which the individual test subjects made while completing the self-tests was related to the comments that students made while going through the materials. All items on the self-tests were scored as being either correct or incorrect and in this case no criterion was applied to the self-test data — all of the raw individual item response data were utilized. The individual items on the self-tests corresponded to specific sections of the instructional materials (a common outcome of instruction which has been designed systematically). Using the comments (identified problems) made during instruction, each self-test item was coded in order to relate test subject performance on self-test items with the perceived problems of the instruction that corresponded to those items. This was done to address the following research question: To what extent do errors made on the post test reflect perceived problems with the instructional materials? The following coding system (reflecting all of the possible kinds of relationships) was developed:

- 1 = a problem was identified with the instruction but the item was responded to correctly
- 2 = a problem was identified with the instruction and the item was responded to incorrectly
- 3 = a problem was not identified with the instruction and the item was responded to incorrectly

4 = a problem was not identified with the instruction and the item was responded to correctly

Frequencies for each kind of relationship were tabulated.

Relating Self-Test Results to Test Subjects' Perceptions of Self-Test Error Causation

During the debriefing procedure, after results of the self-test were made known, test subjects were asked to determine what they perceived might have caused the errors made on the self-tests. For small groups, only errors brought to the agenda were discussed — the criterion was imposed on the self-test item response data. The research question being addressed here is: What are test subjects' perception of self-test error causation? The following coding scheme reflects all of the possible perceived causes of test subjects' making errors on the self-test:

- 1 = there was a problem with the instructional text
- 2 = there was a problem with the self-test item
- 3 = there were problems both with the instructional text and the self-test
- 4 = there were no problems either with the instructional text or the self-test

All subjects within a group were coded with the group's consensus for self-test error causation.

Frequencies for each of the above categories were tabulated based on the debriefed self-test error data.

Relating Comments Made While Taking the Self-Test to Comments Made After Knowledge of Results of the Self-Test

Both problem explication and revision responses (comments) based on debriefing the self-test results (self-test error data) were related to the debriefing outcomes based on discussing the comments which were noted while taking the self-test. The outcomes of having debriefed self-test item error data were coded as either being redundant/verified or not with the outcomes of having debriefed the comments noted while taking the self-test, i.e., comments made about the self-test during instruction.

This coding procedure corresponds with the following research question: After developmental testing, as a debriefing strategy, to what extent does the procedure of having test subjects review their self-test errors produce unique or redundant feedback data? For all test subjects, the individual debriefing outcomes based on discussing the self-test error data were examined in order to compare them with the debriefing outcomes based on reviewing the comments made while taking the self-test. For small groups, there were criteria (as previously discussed) operating both for putting the comments made while taking the self-test on the agenda and for putting the self-test results (errors to be discussed) on the agenda. All subjects within a group were coded with the group's agreed upon response for both problem explications and revisions. Frequencies of redundant responses and unique responses were tabulated.

Relating Comments Based on Knowledge of Attitudinal Survey Results with any Previous Feedback

Responses to attitudinal survey items indicating dissatisfaction were used as a method for identifying problems within the instructional materials and a way to set the debriefing agenda based on survey results. Using the identified problems, based on the attitudinal survey results, one-to-one test subjects were requested to elaborate on why they had indicated dissatisfaction and also how to improve the materials so that another student would not experience similar dissatisfaction. This task resulted in generating a problem explication and revision for each debriefed survey item. As previously discussed, small groups were requested to reach a level of agreement on problem explications and revisions after a group's mean response (or open-ended responses) was used to bring the item to the debriefing agenda. All subjects within a group were coded with the group's agreed upon response for both problem explications and revisions.

For each subject across all experimental conditions, each problem explication and revision based on the results of the attitudinal survey was coded as either being redundant/verified or not with previous feedback gleaned from the other sources of feedback. This addresses the following research question: After developmental testing, as a debriefing strategy, to what extent does the procedure of

having test subjects review their unfavorable attitudinal survey responses produce unique or redundant feedback data? Any of the following feedback sources could have contributed to the redundancy/verification of a problem explication and revision based on survey results:

1. Comments made during instruction (including any comments related to the self-test items)
2. Results from having debriefed the self-test errors
3. Results from another survey question(s)
4. From both comments made during instruction and another survey question(s)
5. Results based on the retrospective debriefing mechanism of going back over comments made during instruction

The above classification system was used to code all debriefed survey item responses by individual subject. Frequencies were tabulated, using the above categories, for both redundant and unique responses.

Different Kinds of Test Subject Comments

In addition, for descriptive purposes, each of the debriefing outcomes related to the comments made during instruction was inspected and coded as to what kind of outcome resulted from the debriefing discussion. This addresses the following research question: Do subjects just locate and not explain problems within the instructional materials or can they explain and/or remediate the problems? The following coding scheme reflects all of the possibilities for purposes of classifying test subjects' responses:

1. The test subject just locates where they perceive a problem to be without explicating or resolving the problem.
2. The test subject just explicates the perceived problem without resolving the problem.
3. The test subject immediately revises/suggests a revision without explication.
4. The test subject locates, explicates, and resolves the problem.

5. There was no previous indication of the problem; the test subject identifies a new problem during debriefing that was undetected during instruction.

Frequencies were tabulated for each classified debriefing outcome.

Content Analysis of Comments: Explicated Problem and Revision Responses

At the level of coding by individual subject, for content analytic purposes, all of the individual problem explications and revisions/revision prescriptions (comments) that were generated by the total sample ($N=72$) during instruction and debriefing were entered verbatim into micro-computer data bases. Subjects had been instructed to write their individual problem explications and revisions directly on the materials themselves (the instruction and accompanying self-test items) during instruction. These comments were included in the content analysis.

During the debriefing, the discussion was recorded by a tape recorder so that the finalized outcomes could be transcribed afterwards. Debriefing produced comments based on three different sources of feedback data: 1) comments made during instruction, 2) self-test error data, and 3) attitudinal survey data. Separate databases were created for all of the problem explications and all of the revisions/revision prescriptions. Separate data bases were maintained for both the developmental testing outcome data (comments on the instruction and self-tests) and the debriefing outcome data. All of the data bases were closely examined for identical or very similar (synonymous) statements. Statements which were identical or very similar were grouped together. The individual statements from the lists of problem explications and revisions/revision prescriptions were then each assigned a code number, e.g., 001, 002, etc. Statements which were identical or very similar were assigned identical code numbers. This resulted in the creation of four master, numbered lists of comments (data).

All of the comments (problem explications and revisions) generated from the study were then content analyzed. The goal of the content analysis was to produce a set of general categories that would

describe the specific kinds of problems and revisions/revision prescriptions in order to account for all of the comments made by the test subjects who participated in the study. This would address the following research question: What general and specific kinds of problem explications and revisions/revision prescriptions result from developmental testing and debriefing?

All of the problem explications and revisions generated through the experimental conditions were coded by using a categorical system developed specifically for this purpose. At first, Debert's (1979, p. 20) summary of the types of revisions that tend to be made in instructional materials after formative evaluation was applied to the outcomes of developmental testing and debriefing sessions. Weston (1986) has noted that the following list of Debert's categories of revision prescriptions is one of the most comprehensive summaries of revisionary tactics:

Addition:

- Add instruction on prerequisite skills and knowledge.
- Add training on the use of the material or the method.
- Add training for teachers or trainers who will use the materials.
- Add a preview (advanced organizer).
- Add illustrations.
- Add performance aids.
- Add examples.
- Add activities.
- Add feedback.
- Add transfer exercises.
- Add test items.
- Add motivation.
- Add variety.

Simplification:

- Reduce the level of complexity.
- Simplify the language.
- Use smaller units.
- Combine materials into larger units.
- Rearrange the sequence.
- Remove irrelevant information.
- Remove irrelevant activities.

Others:

- Make examples more relevant.
- Change the instructional medium.
- Change the instructional design format.
- Change the target population.
- Abort the project.
- Make no changes.

From this list of categories of revision prescriptions, it was apparent that Debert's revision prescriptions tended to take three basic forms: add, delete, or change something within the materials. It was a simple task to generate a parallel list of categories of problem explications in order to complement Debert's original list of categories of revision prescriptions. For example, a complementary problem explication category for the "remove irrelevant information" revision prescription item would be "irrelevant information". The problem is that irrelevant information has been identified — the solution to the problem is the revision prescription to remove the irrelevant information.

In attempting to fit all of the verbatim problem explications and revisions/revision prescriptions into Debert's categories, it was soon discovered that Debert's original categorical system was not

comprehensive enough to classify all of the data (verbatim comments). Debert only dealt with revision prescription whereas the test subjects' comments reflected causation of the problems being encountered. Many of the test subjects' detailed explanations of problems were unique when compared with Debert's specific categories. As a result, a new categorical system was created, based on Debert's original work, which reflected all of the comments generated by the study.

The modification of Debert's categories is a representation of all of the general and specific kinds of problem explications and revisions/ revision prescriptions resulting from the developmental testing and debriefing of the self-instructional materials used in this study. This original representation of test subjects' articulation of learning problems and resolutions addresses the following research question: What general and specific kinds of problem explications and revisions/revision prescriptions result from developmental testing and debriefing? For content analytic purposes, this question was addressed by first applying Debert's original categorical system, using both his general categories and specific subcategories, to the problem explication comments and the revision comments. During the sorting of all of the comments, whenever Debert's classification system was unable to address a given statement, a new subcategory was added to Debert's system. This generally resulted in an expansion of Debert's system, especially the category of "Others".

An inspection of Debert's classification system reveals that three major categories are utilized:

1. Add
2. Simplify
3. Others

A review of all of the categorized comments suggested that they could be reclassified by using more specific kinds of general categories. Expanding the number of general categories better described the specific kinds of comments made by the test subjects. The specific results of this method for conducting the content analysis will be presented in the next chapter.

All of the sub-categories within the modified Debert classification system reflect all of the specific statements (comments) generated through the conduct of this study. These specific statements were closely inspected by two researchers. These researchers reached agreement on the sorting of the

statements into groups which appeared to differ qualitatively. Major categories emerged through the sorting process as similar kinds of problem explications and revisions were being placed together, e.g., all of the problems relating to the format of the materials were grouped together, etc. After grouping all of the statements, six distinctly general or global kinds of problem explications and revisions emerged. These are the major or general categories of the modified Debert classification system:

A System for Classifying Problem Explications and Revisions/Revision Prescriptions

1. Comprehension/Recall kinds of problem explications and revisions/revision prescriptions
2. Motivation/Attention kinds of problem explications and revisions/revision prescriptions
3. Physical Feature kinds of problem explications and revisions/revision prescriptions
4. Multiple Responses overlapping two or more of the above three kinds of problem explications and revisions/revision prescriptions
5. Non-Classifiable Responses for kinds of problem explications and revisions/revision prescriptions
6. Responses indicating that the test subject was unable to explicate a problem and/or suggest a revision after a problem had been located

For all subjects, using these six general categories for both problem explication and revision responses, the original verbatim problem explications and revisions made during instruction were classified into these six general categories for each individual subject, for both the first and second units of instructional material. Frequencies were tabulated, by category, separately for both problem explication and revision/revision prescription responses for all of the subjects (N=72). In addition, the total number of categorized problem explications and revisions for the two units combined were tabulated by general category for each subject. Only problem explication and revision responses which were made during instruction (the comments made about the instruction and the self-tests excluding debriefing response data) were examined, categorized, and tabulated. Tabulating the data in this manner would permit a quantitative analysis of the frequencies (amounts) of the general kinds of comments generated during instruction. This would address the following research questions: Do experimental

conditions affect the amounts of general kinds of problem explications and revisions/revision prescriptions made during instruction? and, Do experimental conditions in conjunction with the test subject characteristic of having indicated prior knowledge of the material affect the amounts of general kinds of problem explications and revisions/revision prescriptions made during instruction?

The next chapter will discuss the data analysis procedures used to address all of the research questions. Results from the analyses will be presented and discussed.

CHAPTER VI

DATA ANALYSIS PROCEDURES AND RESULTS OF THE EXPERIMENTAL MANIPULATION OF DEVELOPMENTAL TESTING VARIABLES

In this chapter the results of both the quantitative and content analyses addressing the various research questions of interest will be presented. CHAPTER VII will be a discussion of these results. Quantitative data analysis procedures were conducted using SPSSX Version 2.1 on an AMDAHL 5801 mainframe computer. Within the data files used in the statistical analysis, data records per subject varied in number. This was due to the fact that different subjects provided varying amounts of data. As a result, a mixed file-type was used in the processing of the data.

Data Analysis Procedures and Results by Research Question

The procedure used to analyze the data related to each research question will be presented first and immediately following this, the results will be presented. All of the research questions were listed on pages 127 and 130

The stimulus materials which were used in the experiment had been produced by teams of experts (subject matter experts, instructional design experts, and evaluation experts) following the guidelines for systematic instructional design. These materials were not contrived for experimental purposes. Any problems inherent within these materials resulted during their development in the field. It is remarkable that the test subjects used in this study identified, explicated, and suggested remediation for as many perceived problems as they did. From an instructional designer's point of view, certainly any kind of perceived problem is worthy of attention.

Research Question #1: Do experimental conditions affect the amount of data obtained from the different sources of feedback? and Do experimental conditions affect the amount of data obtained from the combined sources of feedback?

This first question is general in nature and will be addressed by presenting the results separately for the analysis of each of the four sources of feedback data that were used in the experiment:

Sources: a) *comments made during instruction (including any comments made while responding to individual self-test items)*

b) *debriefed comments*

c) *debriefed self-test errors*

d) *debriefed attitudinal survey results*

Separate two-way univariate analyses of variance were conducted (c.f., Glass and Hopkins, 1984; Keppel, 1982) by the two independent variables: role of participants (with active and passive as the two levels of this independent variable) and number of participants (with one-to-one and small group as the two levels of this independent variable), for the four different sources of feedback data (dependent variables): a) the number of comments (problems being identified) made during instruction (including any comments made while responding to individual self-test items), b) the number of comments that were debriefed, c) the number of debriefed self-test errors, and d) the number of debriefed attitudinal survey responses (unfavorable responses). In addition, the dependent variable of the total number of debriefed agenda items from all sources combined across the two units of instruction was analyzed. Results will be presented by source of feedback. All descriptive statistics related to the results of the two-way ANOVA of the data from the sources can be found in Appendix I. Summary tables and graphic portrayals of the results are incorporated into the text.

a) Comments During Instruction

The frequency of each time a test subject identified a problem was tabulated. Descriptive statistics for this first source of feedback (including any comments made while responding to individual self-test items), are presented in the form of frequency distributions. Tables 11, 12, and 13 (see Appendix I) present descriptive statistics for the total number of problems which were identified (comments).

during instruction by the sample for Unit one, Unit two, and the total number of comments for the two units combined. Tables 14, 15, and 16 present descriptive statistics for the main effect of the variable: role of participants. Tables 17, 18, and 19 present descriptive statistics for the main effect of the variable: number of participants. Tables 20, 21, and 22 present descriptive statistics for the interactions of role and number of participants on the number of problems that were identified during instruction. Table 23 presents a table of means based on the ANOVA results. Table 24 presents the two-way ANOVA summary table of the main effects and interactions of the independent variables on the number of comments made during instruction. Note that separate analyses were conducted for Unit one, Unit two, and for the total (Unit one + Unit two). Figures 10 and 11 present a graphic portrayal of the results of the data analysis. The only significant result observed was the main effect of number of participants, $F(1,68) = 4.68, p < .05$; one-to-one produced significantly more comments than small group for Unit one.

b) Debriefed Comments

Procedures used for the debriefing of comments and the coding of debriefing outcomes can be found in Chapter V.

Tables 25, 26, and 27 (see Appendix I) present descriptive statistics for the total number of debriefing outcomes for the entire sample for Unit one, Unit two, and the total number for the two units combined. Tables 28, 29, and 30 present descriptive statistics for the main effect of the variable role of participants. Tables 31, 32, and 33 present descriptive statistics for the main effect of the variable number of participants. Tables 34, 35, and 36 present descriptive statistics for the interactions of role and number of participants on the number of debriefed comments. Table 37 presents a table of means based on the ANOVA results. Table 38 presents the two-way ANOVA summary table of the main effects and interactions of the independent variables on the number of identified problems (comments) which were debriefed and recorded as debriefing outcomes. Note that separate analyses were conducted for Unit one, Unit two, and for the total (Unit one + Unit two). Figures 12 and 13 present a graphic

Table 24.

Two-Way ANOVA of the Number of Problems Identified (Comments) during Developmental Testing

Unit 1: Source	Sum of squares	df	Mean squares	F	Significance of F
Role of Participants (R)	10.125	1	10.125	1.128	.292
Number of Participants (N)	42.014	1	42.014	4.682	.034 *
R x N Interaction	4.014	1	4.014	.447	.506
Residual	610.167	68	8.973		
Total	666.319	71	9.385		

Unit 2: Source	Sum of squares	df	Mean squares	F	Significance of F
Role of Participants (R)	13.491	1	13.491	1.006	.319
Number of Participants (N)	11.827	1	11.827	0.882	.351
R x N Interaction	46.419	1	46.419	3.462	.067
Residual	898.281	67	13.407		
Total	969.662	70	13.852		

Total: Source	Sum of squares	df	Mean squares	F	Significance of F
Role of Participants (R)	45.825	1	45.825	1.265	.265
Number of Participants (N)	96.267	1	96.267	2.657	.108
R x N Interaction	76.872	1	76.872	2.122	.150
Residual	2427.497	67	36.231		
Total	2644.592	70	37.780		

* $p < .05$

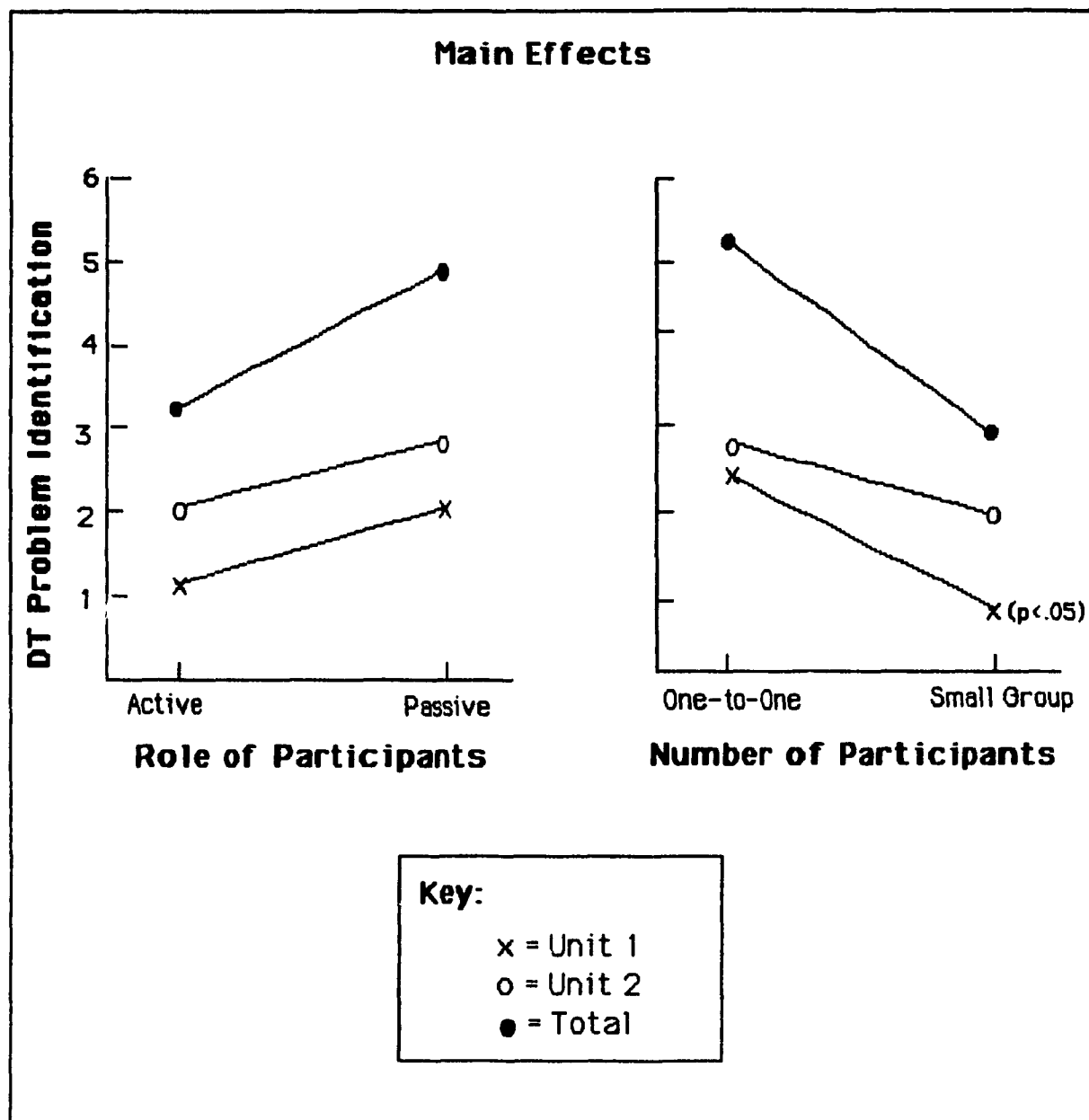


Figure 10. Plotting of means. Main effects of role and number of participants on the number of problems identified by test subjects commenting during the developmental testing of Unit 1, Unit 2, and total for the instructional material.

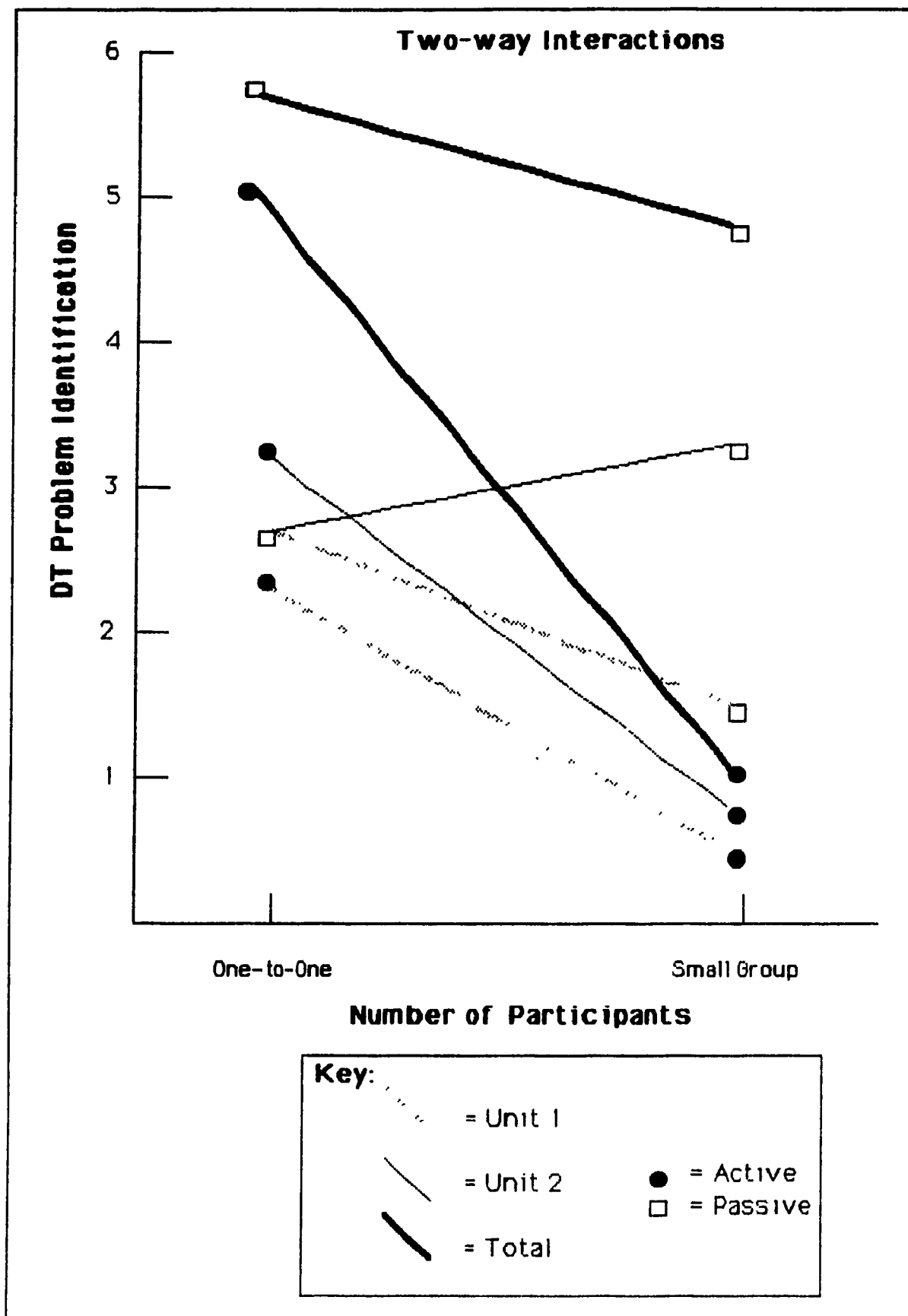


Figure 11 Plotting of means Interaction between role and number of participants with the number of problems identified by test subjects commenting during the developmental testing of Unit 1, Unit 2, and total for the instructional material

Table 38.

Two-Way ANOVA of the Number of Comments that were Debriefed

Unit 1: Source	Sum of squares	df	Mean squares	F	Significance of F
Role of Participants (R)	0.188	1	0.188	0.012	.914
Number of Participants (N)	115.147	1	115.147	7.158	.009**
R x N Interaction	5.475	1	5.475	.340	.562
Residual	1077.778	68	16.086		
Total	1198.479	71	17.121		
Unit 2: Source	Sum of squares	df	Mean squares	F	Significance of F
Role of Participants (R)	21.905	1	21.905	1.855	.178
Number of Participants (N)	0.000	1	0.000	0.000	1.000
R x N Interaction	22.549	1	22.549	1.910	.172
Residual	791.007	67	11.806		
Total	835.465	70	11.935		
Total: Source	Sum of squares	df	Mean squares	F	Significance of F
Role of Participants (R)	32.000	1	32.000	0.866	.355
Number of Participants (N)	128.000	1	128.000	3.462	.067
R x N Interaction	50.000	1	50.000	1.352	.249
Residual	2514.000	68	36.971		
Total	2724.000	71	38.366		

** $p < .01$

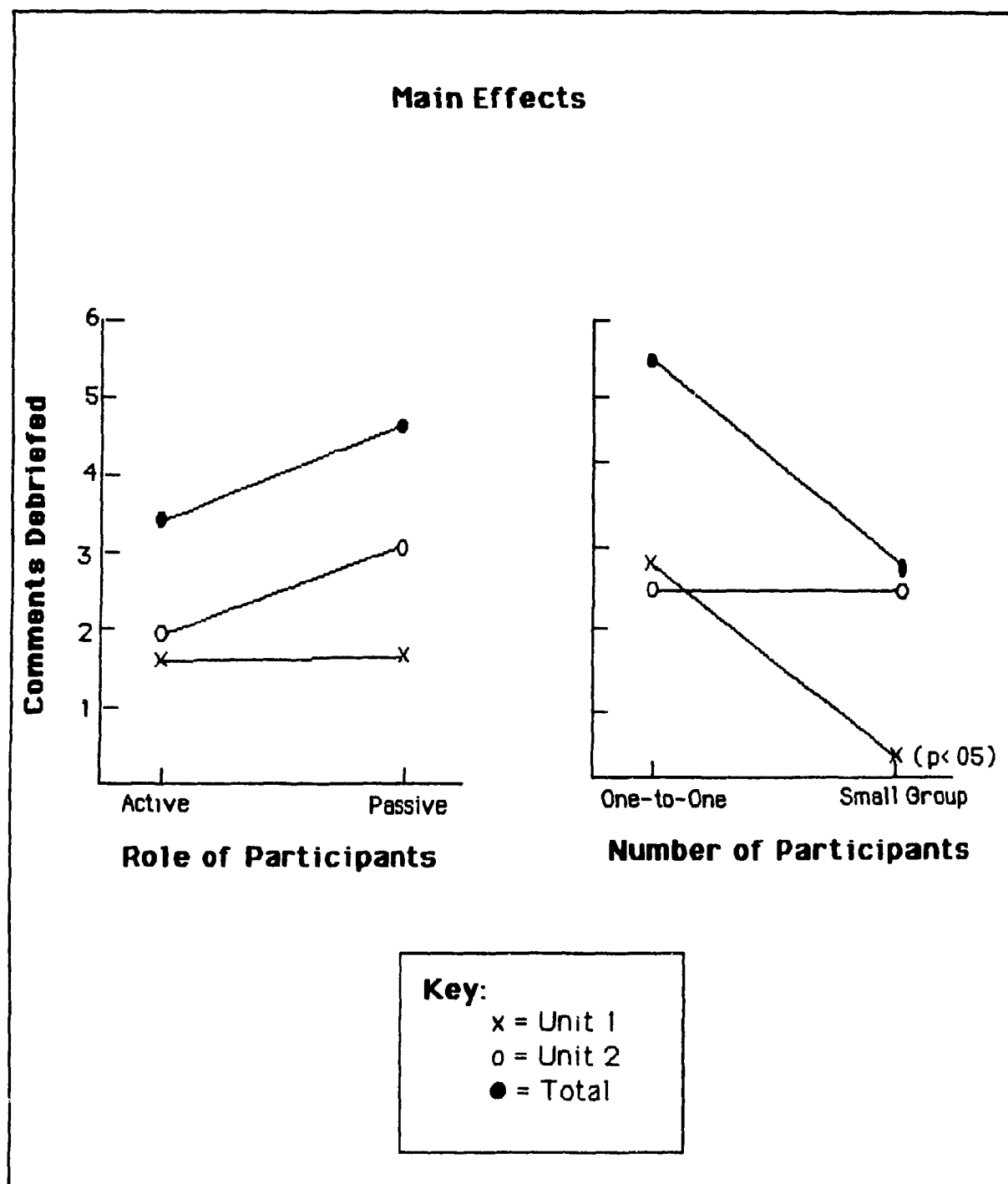


Figure 12. Plotting of means: Main effects of role and number of participants on the number of debriefing agenda items based on comments for Unit 1, Unit 2, and total for the instructional material.

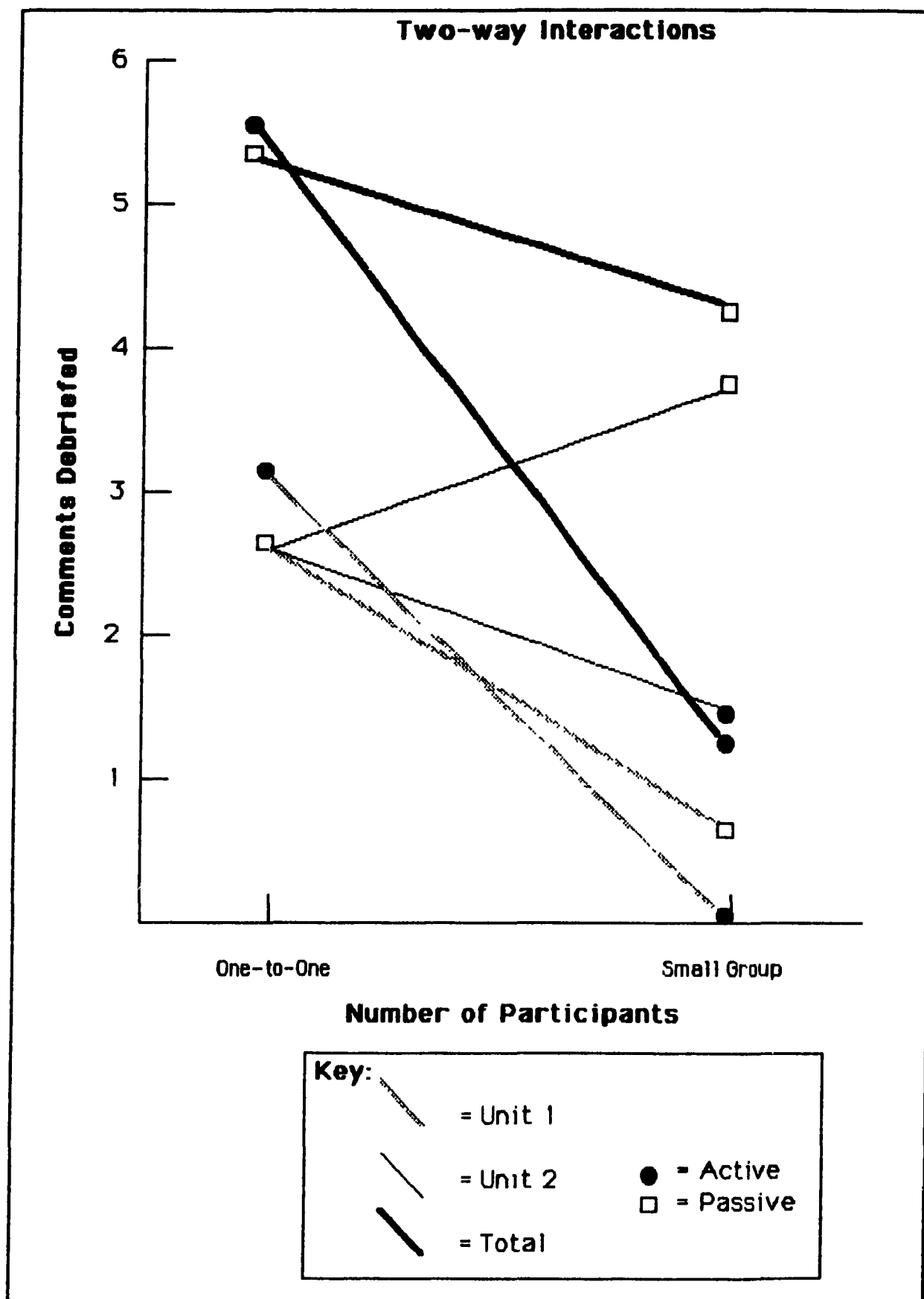


Figure 13. Plotting of means. Interaction between role and number of participants with the number of debriefing agenda items based on comments for Unit 1, Unit 2, and total for the instructional material.

portrayal of the results of the data analysis. The only significant result obtained was the main effect of number of participants for the first unit, $F(1,68) = 7.16, p < .01$.

c) Debriefed Self-Test Errors

The next source of feedback to be analyzed was the number of debriefing outcomes based on self-test errors. Procedures used for the debriefing of self-test errors and the coding of error data can be found in Chapter V.

Tables 39, 40, and 41 (see Appendix I) present descriptive statistics for the total number of self-test errors which were debriefed by the total sample for Unit one, Unit two, and the total number for the two units combined. Tables 42, 43, and 44 present descriptive statistics for the main effect of the variable role of participants. Tables 45, 46, and 47 present descriptive statistics for the main effect of the variable number of participants. Tables 48, 49, and 50 present descriptive statistics for the interactions of role and number of participants on the number of self-test errors that were debriefed. Table 51 presents a table of means based on the ANCOVA results. Table 52 presents the two-way ANOVA summary table of the main effects and interactions of the independent variables on the number of self-test errors which were debriefed. Note that separate analyses were conducted for Unit one, Unit two, and for the total (Unit one + Unit two). Figures 14 and 15 present a graphic portrayal of the results of the data analysis. A significant main effect was observed for the number of participants for Unit one, $F(1,68) = 18.124, p < .0001$, for Unit two, $F(1,67) = 14.079, p < .0001$, and for Units one and two combined, $F(1,68) = 20.213, p < .0001$. A significant interaction between role and number of participants was found for Unit one, $F(1,68) = 12.44, p < .001$, for Unit two, $F(1,67) = 4.01, p < .05$, and for Units one and two combined, $F(1,68) = 11.274, p < .001$.

d) Attitude Survey

The next source of feedback to be analyzed was the number of debriefing outcomes based on unfavorable attitudinal survey responses. Procedures used for the debriefing of unfavourable responses and the attitudinal data can be found in Chapter V.

Table 52.

Two-Way ANOVA of the Number of Debriefing Agenda Items Based on Posttest Errors

Unit 1: Source	Sum of squares	df	Mean squares	F	Significance of F
Role of Participants (R)	3.125	1	3.125	3.329	.072
Number of Participants (N)	17.014	1	17.014	18.124	.0001****
R x N Interaction	11.681	1	11.681	12.443	.001 ***
Residual	63.833	68	.939		
Total	95.653	71	1.347		

Unit 2: Source	Sum of squares	df	Mean squares	F	Significance of F
Role of Participants (R)	0.415	1	0.415	0.182	.671
Number of Participants (N)	32.194	1	32.194	14.079	.0001****
R x N Interaction	9.169	1	9.169	4.010	.049 *
Residual	153.209	67	2.287		
Total	195.099	70	2.787		

Total: Source	Sum of squares	df	Mean squares	F	Significance of F
Role of Participants (R)	8.681	1	8.681	2.024	.159
Number of Participants (N)	86.681	1	86.681	20.213	.0001****
R x N Interaction	48.347	1	48.347	11.274	.001 ***
Residual	291.611	68	4.288		
Total	435.319	71	6.131		

* $p < .05$ *** $p < .001$ **** $p < .0001$

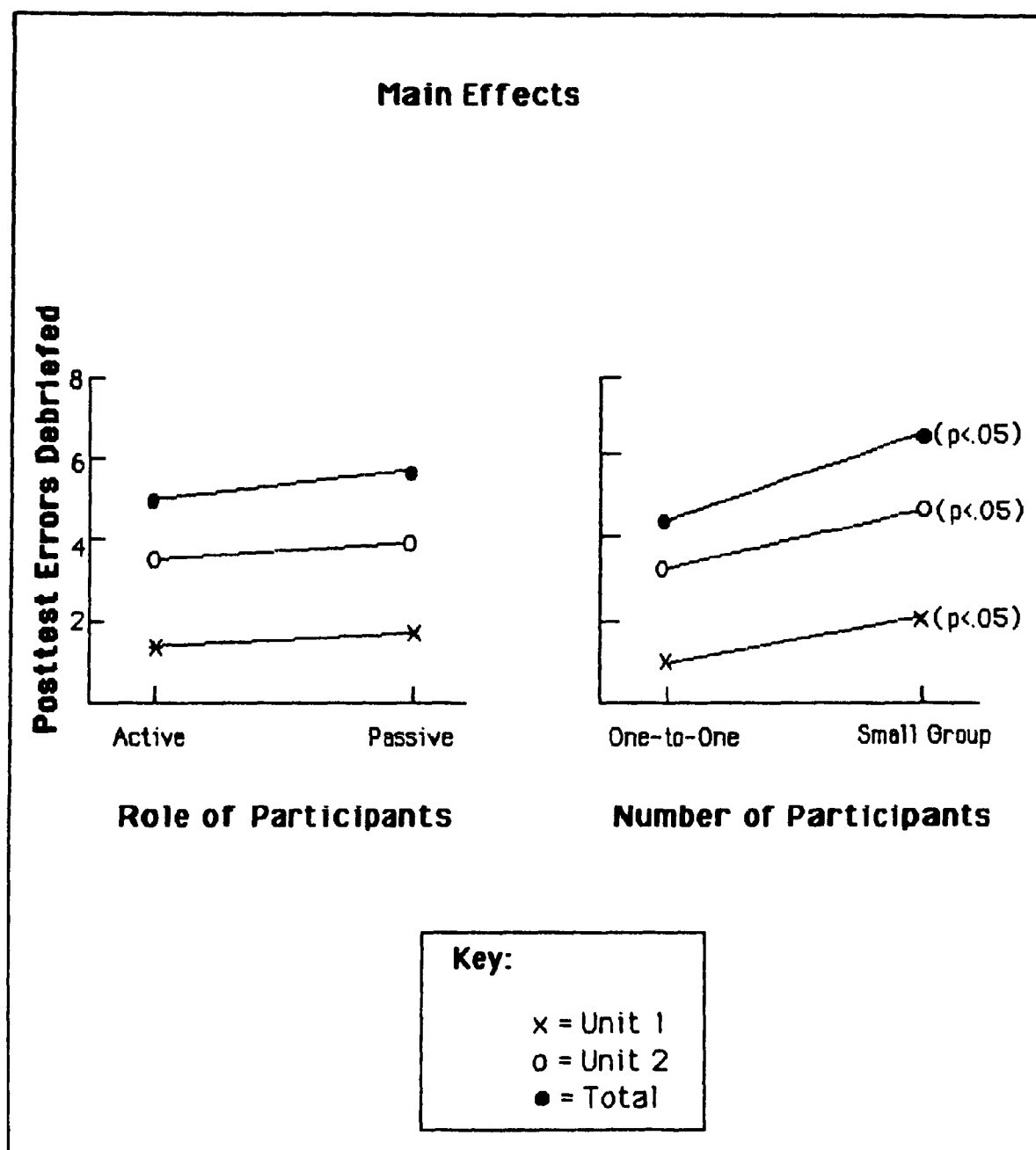


Figure 14. Plotting of means. Main effects of role and number of participants on the number of debriefing agenda items based on posttest errors for Unit 1, Unit 2, and total for the instructional material.

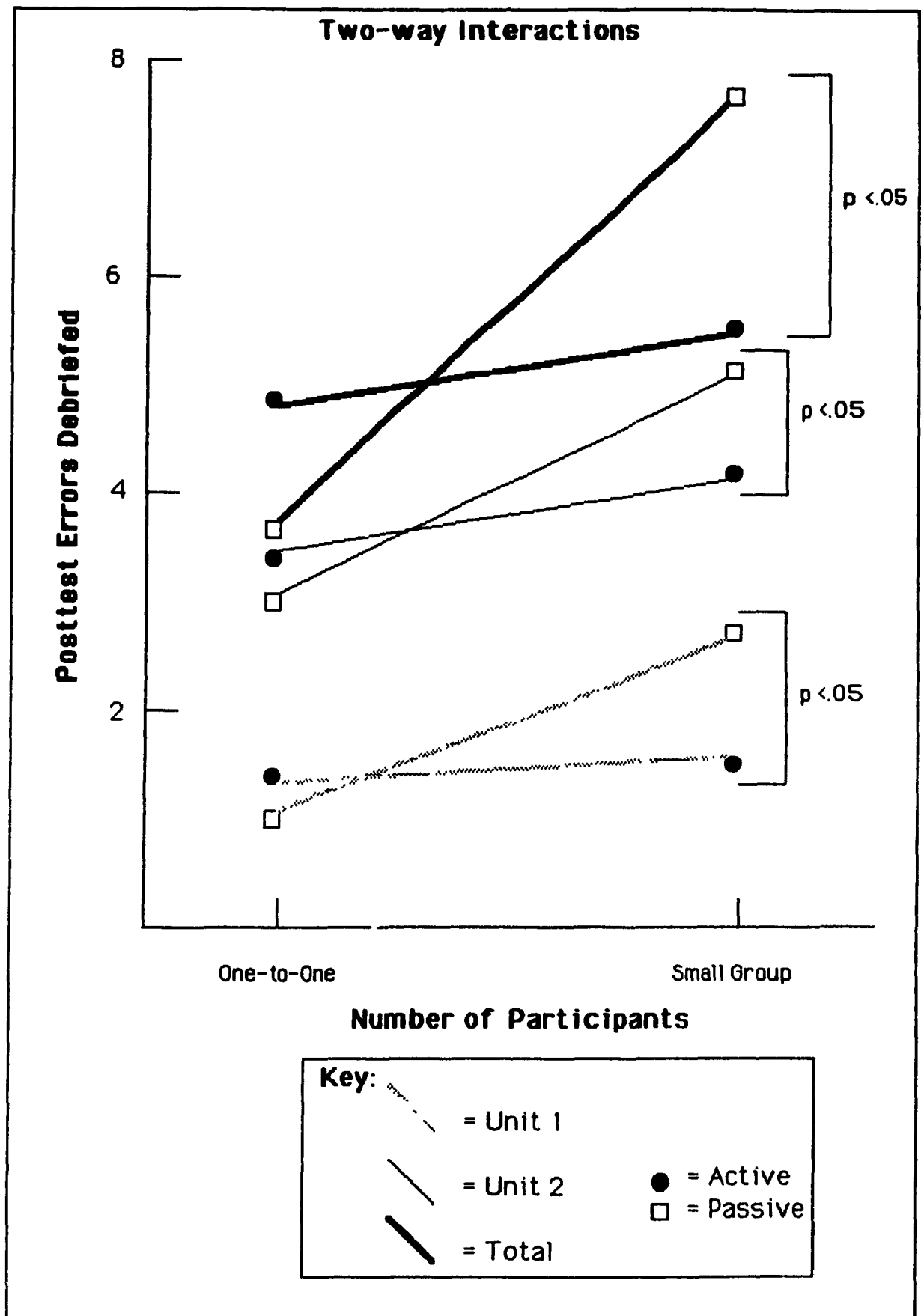


Figure 15. Plotting of means. Interaction between role and number of participants with the number of debriefing agenda items based on posttest errors for Unit 1, Unit 2, and total for the instructional material.

Tables 53, 54, and 55 (see Appendix I) present descriptive statistics for the total number of debriefing outcomes (problems identified) based on unfavorable attitudinal survey responses by the sample for Unit one, Unit two, and the total number for the two units combined. Tables 56, 57, and 58 present descriptive statistics for the main effect of the variable role of participants. Tables 59, 60, and 61 present descriptive statistics for the main effect of the variable number of participants. Tables 62, 63, and 64 present descriptive statistics for the interactions of role and number of participants on the number of debriefing agenda items based on unfavorable attitudinal survey responses. Table 65 presents a table of means based on the ANOVA results. Table 66 presents the two-way ANOVA summary table of the main effects and interactions of the independent variables on the number of unfavorable attitudinal survey items which were debriefed. Note that separate analyses were conducted for Unit one, Unit two, and for the total (Unit one + Unit two). Figures 16 and 17 present a graphic portrayal of the results of the data analysis.

Results of the data analysis found no significant interactions but significant main effects for Unit two were revealed. With Unit two, both the main effects of participants' role, $F(1,67) = 7.26, p < .01$, and number of participants, $F(1,67) = 7.87, p < .01$, were statistically significant. No significant differences were detected for Unit one or for both units combined.

Do experimental conditions affect the amount of data obtained from the combined sources of feedback?

The following analysis is not related to one source of feedback but uses an aggregate of all of the various sources of feedback summed across the two units. Using the total number of all of the finalized debriefing outcomes, an index was created in the form of an aggregated score representing the total amount of feedback data gathered from a test subject. Table 67 (see Appendix I) presents descriptive statistics for the total number of debriefed agenda items from all sources combined (comments, self-test results, and attitudinal survey results) across the two units. Table 68 presents descriptive statistics for the main effect of the variable role of participants. Table 69 presents descriptive statistics for the main effect of the variable number of participants. Table 70 presents descriptive statistics for the interactions of the independent variables on the total number of debriefing agenda items from all sources combined. Table 71 presents the two-way ANOVA summary table of the main

Table 66.

Two-Way ANOVA of the Number of Debriefing Agenda Items Based on the Attitudinal Survey Responses

Unit 1: Source	Sum of squares	df	Mean squares	F	Significance of F
Role of Participants (R)	18.000	1	18.000	0.897	.347
Number of Participants (N)	2.000	1	2.000	.100	.753
R x N Interaction	43.556	1	43.556	2.170	.145
Residual	1364.889	68	20.072		
Total	1428.444	71	20.119		

Unit 2: Source	Sum of squares	df	Mean squares	F	Significance of F
Role of Participants (R)	136.746	1	136.746	7.258	.009**
Number of Participants (N)	148.241	1	148.241	7.869	.007**
R x N Interaction	50.637	1	50.637	2.688	.106
Residual	1262.248	67	18.840		
Total	1602.000	70	22.886		

Total: Source	Sum of squares	df	Mean squares	F	Significance of F
Role of Participants (R)	60.500	1	60.500	1.175	.282
Number of Participants (N)	180.500	1	180.500	3.505	.065
R x N Interaction	0.500	1	0.500	0.010	.922
Residual	3502.000	68	51.500		
Total	3743.500	71	52.725		

** $p < .01$

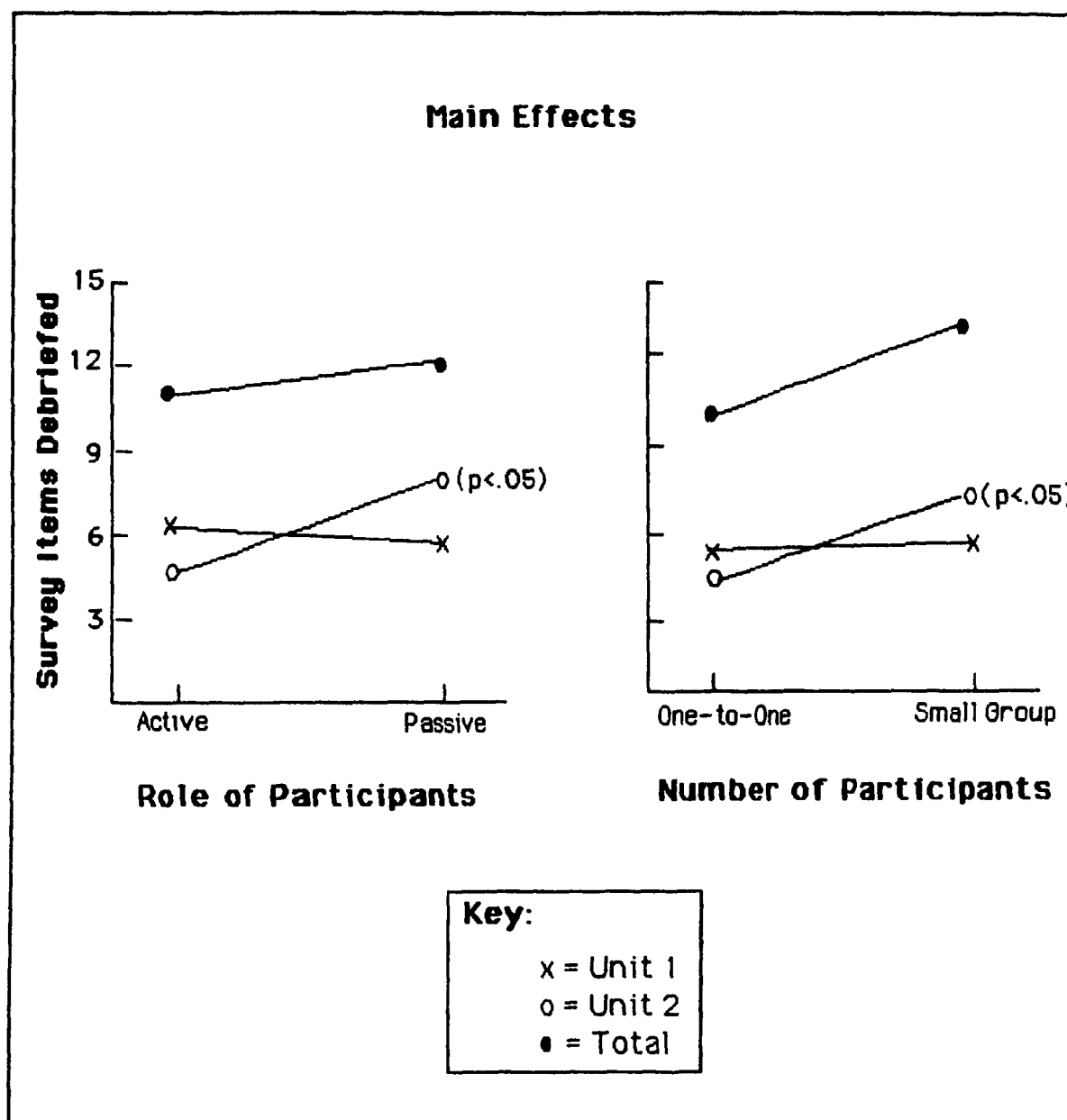


Figure 16. Plotting of means. Main effects of role and number of participants on the number of debriefing agenda items based on survey responses to Unit 1, Unit 2, and total for the instructional material.

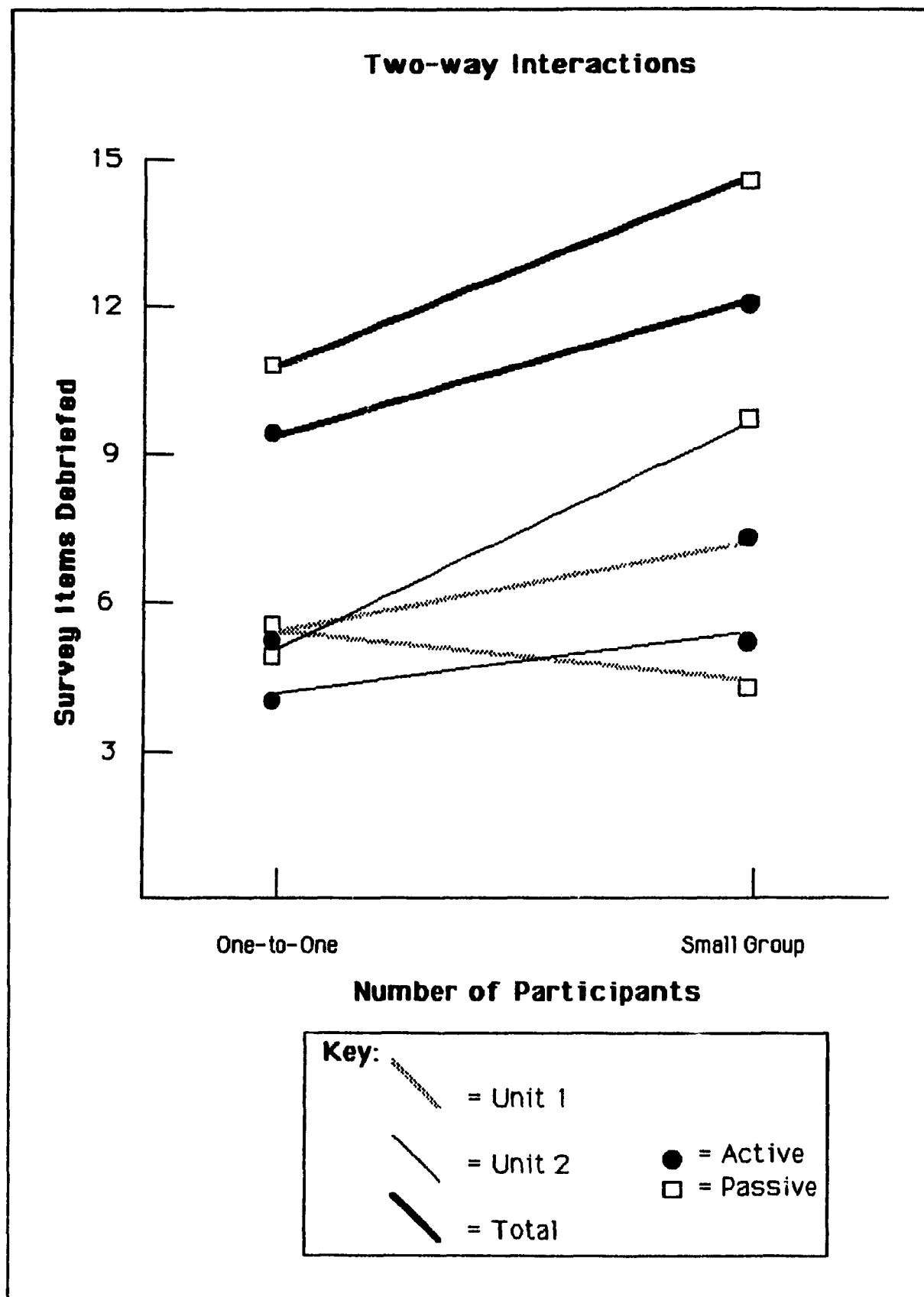


Figure 17. Plotting of means. Interaction between role and number of participants with the number of debriefing agenda items based on survey responses to Unit 1, Unit 2, and total for the instructional material.

Table 71.

Two-Way ANOVA of the Total Number of Debriefing Agenda Items from all Sources Combined

TOTAL: Source	Sum of squares	df	Mean squares	F	Significance of F
Role of Participants (R)	234.722	1	234.722	1.792	.185
Number of Participants (N)	122.722	1	122.722	0.937	.336
R x N Interaction	206.722	1	206.722	1.579	.213
Residual	8905.111	68	130.958		
Total	9469.278	71	133.370		

effects and interactions of the independent variables on the total number of debriefed agenda items from all sources. Figure 18 presents a graphic portrayal of the results of the data analysis. No significant differences were found.

Research Question #2: Does the test subject characteristic of having indicated prior knowledge of the material affect the overall amount of comments being made during instruction?

For the second research question, prior knowledge was used as the independent variable. Two levels of prior knowledge were utilized in the statistical analyses: 1) test subjects had indicated that they had prior knowledge of the instructional materials, and 2) test subjects had indicated that they had no prior knowledge of the instructional materials. The t-test was applied separately to the following dependent variables: 1) the number of comments (problems identified) made during instruction for Unit one, 2) the number of comments (problems identified) made during instruction for Unit two, and 3) the number of comments (problems identified) made during instruction for the two units combined, i.e., the total number of comments. Using prior knowledge (yes or no) as the independent variable, t-tests were separately applied to the number of comments (problems identified) for Unit one, Unit two, and for the total number of comments. Two-tailed probability values are reported. For Unit one, the t-test did not reveal a significant difference between the two groups, $t = -1.32$, $df = 69$, $p > .05$. For Unit two, the t-test did reveal a significant difference between the two groups, $t = -2.29$, $df = 69$, $p < .05$. Those test subjects who had indicated that they had prior knowledge of the material produced more comments than those who did not indicate having prior knowledge. For the two units combined, the T-test did reveal a significant difference between the two groups, $t = -2.05$, $df = 69$, $p < .05$. Again, those indicating prior knowledge commented more frequently than those not indicating prior knowledge.

Research Question #3 To what extent do data from the various sources of feedback provide redundant or unique information? This is a general question addressed by more specific questions:

Specifically: a) After instruction, as a debriefing strategy, to what extent does the procedure of having test subjects go over their comments about the

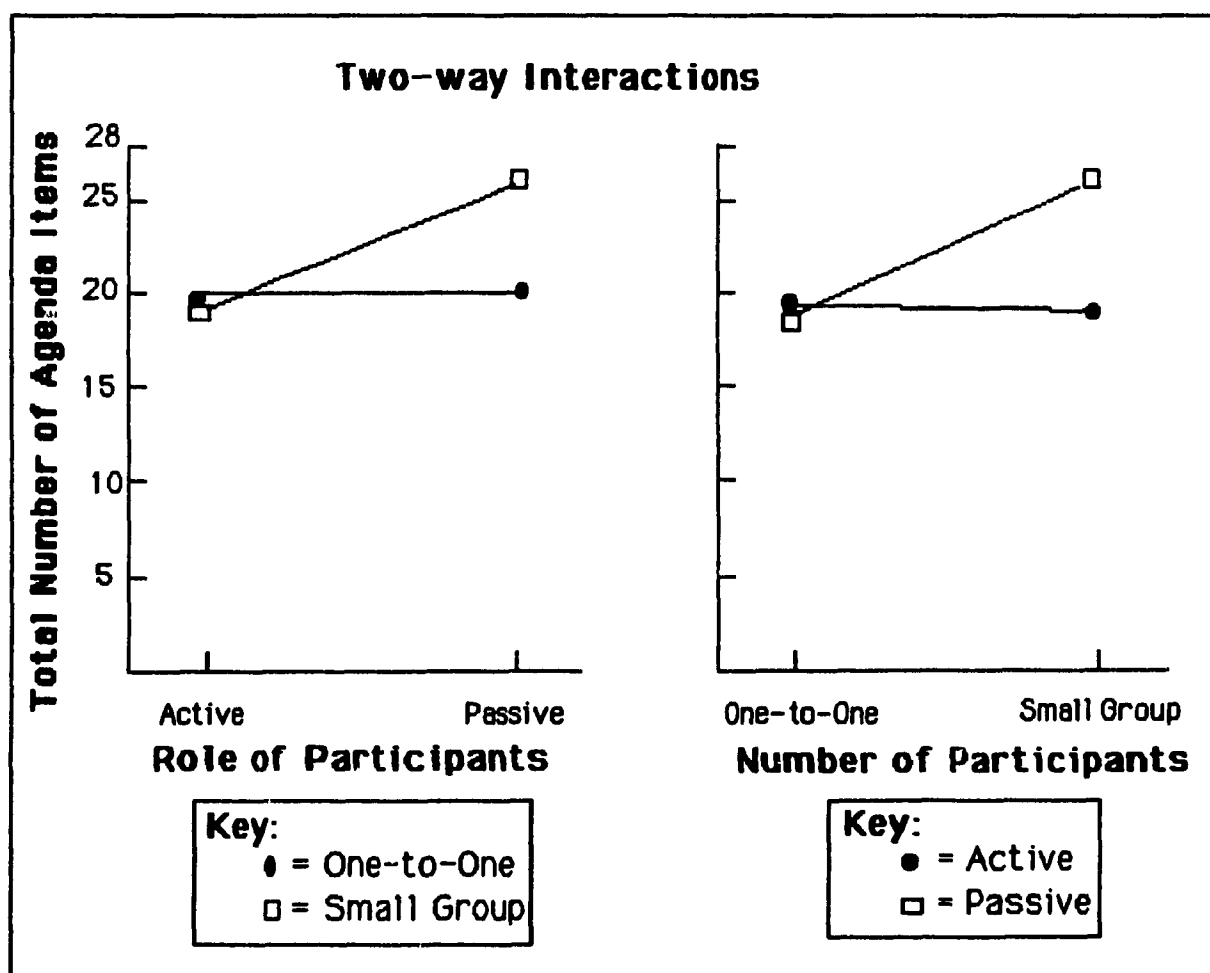
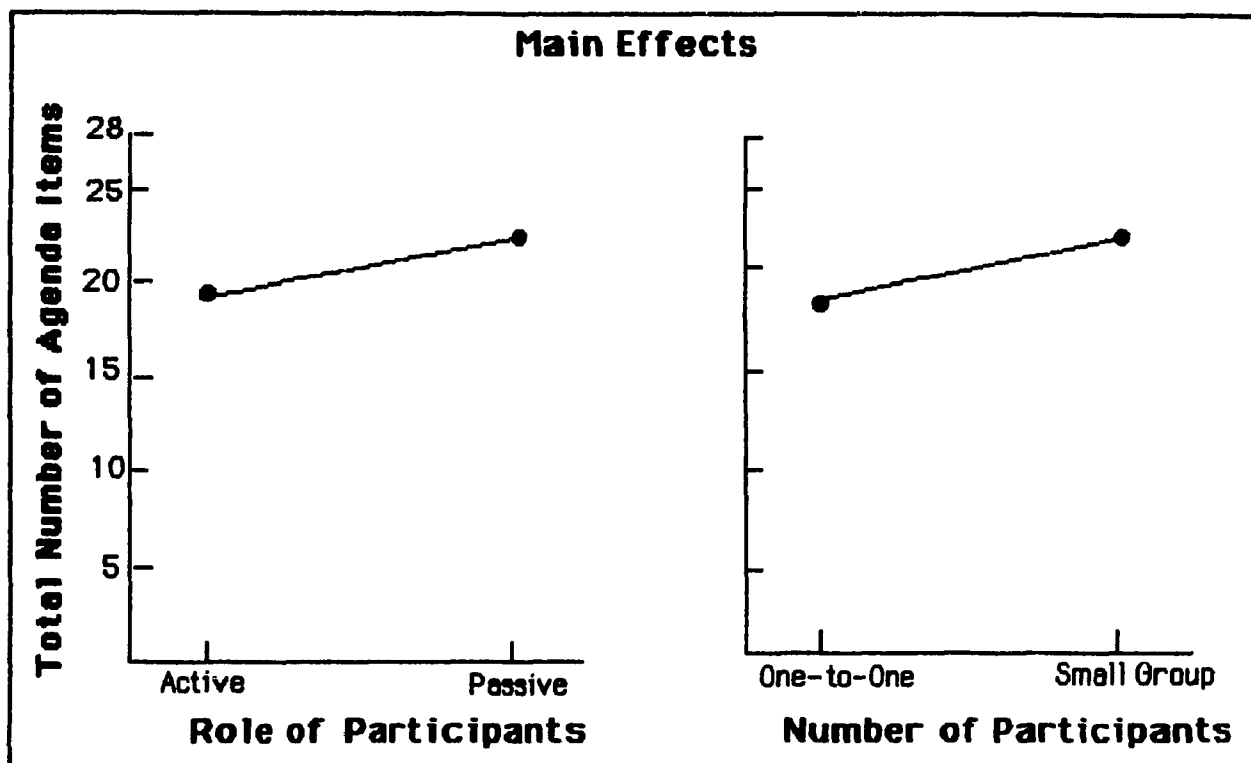


Figure 18. Plotting of means. Main effects of role and number of participants and their interactions with the total number of agenda items (from all feedback sources) debriefed for both units of instructional material.

materials again (in retrospect) produce unique or redundant feedback data?

- b) To what extent does performance on individual self-test items reflect comments (perceived problems) made during instruction?*
- c) What are test subjects' perceptions of self-test error causation?*
- d) After instruction, as a debriefing strategy, to what extent does the procedure of having test subjects review their self-test errors produce unique or redundant feedback data?*
- e) After instruction, as a debriefing strategy, to what extent does the procedure of having test subjects review their unfavorable attitudinal survey responses produce unique or redundant feedback data?*

a) Unique or Redundant Feedback Before and During Debriefing

To answer the first sub-question, all of the outcomes from having debriefed the comments were compared to the original agenda items reflecting comments made during instruction, i.e., the input for debriefing. Debriefing outcomes were coded as either being redundant/verified or not with the comments made during instruction in order to relate the comments made during instruction to the comments made after instruction. The criterion applied to small group comments made during instruction for purposes of determining the input for debriefing should be recalled (see pages 123 - 124). No such criterion was applied to one-to-one comments.

Frequencies and percentages were computed separately, by individual subject ($N=72$) as the unit of analysis, for both problem explication responses and revision responses. Experimental conditions were ignored for this data analysis procedure. It should be recalled that in all experimental conditions, with respects to debriefing test subject comments, subjects were permitted to both identify problems that had not been noticed during instruction and also free to change their minds — what was a problem during instruction might not have been perceived as problematic upon debriefing. During debriefing test subjects were at liberty to add new comments, delete previous comments, verify previous comments, and also modify previous comments.

Results of the data analysis indicate that for the 321 problem explications resulting from the discussions during debriefings, 51.4% (165/321) of the problem explication debriefing outcomes were non-redundant/unverified with the original input for debriefing. For the 321 revisions/revision

prescriptions resulting from the discussions during debriefings, 55.5% (178/321) of the revision outcomes were non-redundant/unverified with the original input for debriefing.

b) Unique or Redundant Feedback: Self-Test Results and Comments During Instruction

Frequencies were tabulated for each test subject's response to self-test items. In this case, no criterion was applied to the results of the self-test; all of the raw self-test data were used. Comments made about the self-test items during instruction were ignored for this analysis as the intent was to relate performance on a given self-test item to the comments made by subjects directed towards the specific section of the instructional materials which presented the subject matter content for the self-test item. The set of categories presented below reflects all of the possible relationships between performance on self-test items and comments (identified problems) made during instruction. All of the individual self-test items were keyed to the particular sections of the instructional material which provided instruction for successful test performance. As comments were made by section of the instruction, this enabled the investigation of the relationship between commenting and test item performance.

Frequencies and percentages were computed for each category reflecting all of the responses to self-test items that were made by individual subjects (N=72), while ignoring experimental conditions. The results are summarized below:

<u>Category:</u>	<u>Frequency of:</u>	<u>Percent</u>
1. A problem was identified with the instruction but the related self-test item was responded to correctly.	69	5.33
2. A problem was identified with the instruction and the related item was responded to incorrectly.	45	3.47
3. No problem was identified with the instruction and the related item was responded to incorrectly.	274	21.16
4. No problem was identified with the instruction and the related item was responded to correctly.	907	70.04
	<u>1295</u>	<u>100.00</u>

c) Subjects' Perceptions of Causation of Self-Test Errors

Frequencies and percentages were computed for all of the self-test errors that were debriefed. During the debriefing of self-test errors, test subjects were asked to determine what might have been the cause of making the error. The previous question revealed that many of the errors made on the self-test items did not have associated problems identified during instruction. The present research question investigated test subjects' perceptions of problems with the instruction that were not detected while going through the text. In retrospect, after they have been informed of self-test item failure, are perceived problems identified?

For the analysis, experimental conditions were ignored and the individual test subject's (N=72) debriefing agenda, reflecting debriefed self-test errors, was used as the level of analysis. The categories in the following table represent all of the possible relationships between an error on the self-test which was debriefed and test subjects' perceptions of what caused the error.

The results of tabulating the frequencies for each category are summarized below:

<u>Category:</u>	<u>Frequency of:</u>	<u>Percent</u>
1. There was a problem with the instructional text.	138	35.48
2. There was a problem with the self-test item.	53	13.62
3. There were problems both with the instructional text and the self-test.	13	3.34
4. There were no problems either with the instructional text or the self-test.	185	47.56
	389	100.00

In retrospect, 35.48% of the responses did find a problem that had not been spotted during instruction. This procedure certainly added more developmental testing data.

d) Unique or Redundant Feedback: Debriefing Comments on Problems Noted on the Self-Test

Redundant feedback with the results of having debriefed the self-test error data could feasibly have been generated from having discussed, after instruction, the written comments that test subjects made about the individual self-test items while writing the self-test. As a source of feedback, the comments made about the test items during instruction and discussed during the debriefing were temporally unrelated to the discussion of the self-test error data because up to that point in the experimental procedure the items had not been scored. For all experimental conditions, self-test error data were discussed in debriefing sessions immediately after the written comments directed towards the self-test items had been discussed. When it came time to debrief the written comments made about the self-test items during instruction test subjects did not have knowledge of the results of their self-test performance.

Experimental conditions were ignored for this analysis and each one of the individual test subject's (N=72) debriefing outcomes, reflecting discussed self-test errors, was compared with the debriefing outcomes related to the discussion of comments made during instruction while writing the self-test. The outcome from having discussed each self-test error was coded as either being redundant or not with debriefing outcomes based on having discussed the written comments made about the self-test while test subjects were writing the self-test, i.e., during instruction. It should be recalled that comments refer to both problem explication and revision/revision prescription responses. All subjects within a group were coded with the group's agreed upon response for both problem explications and revisions. Frequencies and percentages were computed for both redundant and non-redundant problem explications as well as redundant and non-redundant revisions/revision prescriptions. A total number of 389 self-test errors were debriefed with the sample (N=72). Results of the data analysis indicate that for debriefed self-test item errors only 7/389 or 1.8% of the problem explication debriefing outcomes based on this source of feedback were redundant/verified with the debriefed comments made about the self-test while writing the self-test. Unique feedback was obtained 98.2% of the time (382/389). Only 6/389 or 1.54% of the revision debriefing outcomes related to self-test item error data were

verified/redundant with the debriefed comments made about the self-test while writing the self-test. Unique feedback was obtained 98.46% of the time (383/389).

e) Review of Comments on Survey Responses

All of the debriefed comments (individual problem explications and revisions) generated based on having test subjects review their unfavorable attitudinal survey responses were coded as either being redundant/verified or not with comments from any previous source(s) of feedback. Frequencies and percentages were computed for each debriefing outcome based on discussing the results of the attitudinal survey, by individual subject (N=72) as the unit of analysis. Experimental conditions were ignored for this data analysis procedure. As previously discussed, there was a criterion applied to small group attitudinal survey response data for agenda setting and debriefing purposes. Small groups were requested to reach a level of agreement on problem explications and revisions after a group's mean response (or open-ended response) was used to bring the item to the debriefing agenda. All subjects within a group were coded with the group's agreed upon response for both problem explications and revisions.

From transcripts of all of the debriefing outcomes, 813 problem explications (debriefing comments) were closely inspected and categorized as to whether this feedback data had already been gathered from any previous source of feedback. All of the 813 problem explications were categorized as follows based on the data analysis:

<u>Source of Redundancy/Verification</u>	<u>Frequency</u>	<u>Percent</u>
1. Non-redundant problem explications	197	24.2
2. From comments made during instruction	68	8.4
3. From debriefing the self-test errors	183	22.5
4. From another survey question(s)	54	6.6
5. From both comments made during instruction and another survey question(s)	21	2.6
6. From the retrospective mechanism of going back over comments made during instruction (debriefing)	34	4.2
7. Test subject identifies but can not explicate the problem	11	1.4
8. An unfavourable response to a survey item but discussion of the item response during debriefing indicates that there was no problem	245	30.1
	813	100.0

Results of the data analysis indicate that for problems identified and explained during the debriefing of unfavorable attitudinal survey responses, 44.28% (360/813) of these problem explication debriefing outcomes based on the results of the attitudinal survey were verified/ redundant with the problem explications that test subjects made based on any previous source of feedback. Unique feedback was obtained 24.23% of the time (197/813). It should be noted that categories seven and eight did not result in revision statements.

The experimental procedure required test subjects to suggest a way to revise the material in order to remedy the identified problem. As for problem statements, there were 813 revisions/revision prescriptions to complement the identified problems. Similar to the problem explications, all of the 813 revisions/revision prescriptions (debriefing comments) were categorized as follows based on the data analysis:

<u>Source of Redundancy/Verification</u>	<u>Frequency</u>	<u>Percent</u>
1. Non-redundant revision responses	211	26.00
2. From comments made during instruction	51	6.30
3. From debriefing the self-test errors	122	15.00
4. From another survey question(s)	100	12.30
5. From both comments made during instruction and another survey question(s)	21	2.60
6. From the retrospective mechanism of going back over comments made during instruction (debriefing)	35	4.30
7. Test subject can not make or suggest a revision	26	3.20
8. An unfavourable response to a survey item but discussion of the item response during debriefing indicates that there was no need to revise	247	30.40
	813	100.00

Results of the data analysis indicated that for revision, 40.47% (329/813) of the revision debriefing outcomes based on the results of the attitudinal survey were verified/redundant with the revisions that test subjects made based on any previous source of feedback. Unique feedback was obtained 26% of the time (211/813). Again, it should be noted that categories seven and eight did not result in revision statements.

Research Question #4: Do subjects just locate and not explain problems within the instructional materials or can they explain and/or remediate the problems?

For this question, each of the debriefing outcomes related to the comments made during instruction was inspected and coded as to what kind of outcome resulted from the debriefing discussion. All of the individual responses (individual problem explication and revision outcomes) based on debriefing the

comments which were made during instruction were coded as indicated in the table below. Frequencies and percentages were computed, by individual subject (N=72) as the unit of analysis, for each of the categories. Experimental conditions were ignored for this data analysis procedure. The results of tabulating the frequencies by category are summarized below:

<u>Category:</u>	<u>Frequency of:</u>	<u>Percent</u>
1. The test subject just locates where they perceive a problem to be without explicating or resolving the problem.	50	16.78
2. The test subject just explicates the perceived problem without resolving the problem.	70	23.49
3. The test subject immediately resolves the problem without explication.	55	18.46
4. The test subject locates, explicates, and resolves the problem.	118	39.60
5. There was no previous indication of the problem; the test subject identifies a new problem during debriefing that was undetected during instruction.	5	1.68
	298	100.00

Research Question #5: What general and specific kinds of problem explications and revisions/ revision prescriptions result from developmental testing and debriefing?

This question was addressed by synthesizing the problem explication and revision data bases to the point where only mutually exclusive statements remained in the data bases — all similar and identical statements were reduced to statements that differed qualitatively. Then, Debert's original categorical system was applied to these pools of statements for content analytic purposes. Experimental conditions were ignored during the content analysis. Whenever Debert's classification system was unable to

address a given statement, a new sub-category was added. The results of this method for conducting the content analysis is represented in Tables 72 and 73. Table 72 presents the modified Debert categorical system representing all of the problem explication responses (comments), from all sources of feedback, which were generated from the study.

Table 72.

General Categories and Sub-categories for Problem Explication Responses

1. COMPREHENSION/RECALL KINDS OF PROBLEM EXPLICATIONS:

- 101 = needs additional instruction on how to use materials
- 102 = an advance organizer/preview is needed
- 103 = needs more illustration
- 104 = an example(s) needs to be added
- 105 = needs a summary at the end
- 106 = sentences are too long
- 107 = chunking problem - too much information to remember
- 108 = sequence problem
- 109 = unnecessary information
- 110 = example needs to be more relevant
- 111 = title does not relate to the content
- 112 = a term(s) needs to be defined
- 113 = syntax problem
- 114 = discrimination problem
- 115 = explanation needs to be added
- 116 = comprehension problem - I do not understand
- 117 = explanation needs to be added + add examples
- 118 = the relationship between content and objectives is not clear
- 119 = the list of objectives is incomplete
- 120 = explanation is too general/vague
- 121 = needs to be more explicit
- 122 = it is not clear as to how many responses are required
- 123 = needs better transition between sections
- 124 = the question is not specific enough
- 125 = needs more illustration + a term(s) needs to be defined
- 126 = content does not relate to section heading

2. MOTIVATION/ATTENTION KINDS OF PROBLEM EXPLICATIONS:

- 201 = material is boring
- 202 = irrelevant information
- 203 = the level of complexity needs to be increased
- 204 = test items need to be added
- 205 = example(s) needs to be more relevant
- 206 = redundant information

- 207 = the level of complexity needs to be increased + add test items
- 208 = content needs to be related more to nursing
- 209 = needs more illustration

3. PHYSICAL FEATURE KINDS OF PROBLEM EXPLICATIONS:

- 301 = typeface is annoying
- 302 = the black lines are annoying
- 303 = format/layout problem
- 304 = lower case is poor for purposes of highlighting
- 305 = key concepts/terms need to be underlined
- 306 = sections need to be numbered
- 307 = title is not prominent
- 308 = typographic error
- 309 = format disrupts topicalization
- 310 = spacing is not proportional + the black lines are annoying
- 311 = inconsistent use of underlining for highlighting purposes
- 312 = the main points do not stand out (refers to the need for underlining)
- 313 = other media need to be available
- 314 = format is inconsistent

4. MULTIPLE RESPONSES OVERLAPPING TWO OR MORE OF THE ABOVE CATEGORIES FOR KINDS OF PROBLEM EXPLICATIONS:

- 401 = lower case is poor for purposes of highlighting + add explanation
- 402 = redundant + comprehension problem (I do not understand)
- 403 = format disrupts topicalization + an advance organizer/preview is needed
- 404 = redundant information + discrimination problem
- 405 = typographic error + an explanation is needed
- 406 = syntax problem + redundant information + an explanation needs to be added
- 407 = redundant information + the level of complexity needs to be increased
- 408 = redundant information + typographic error
- 409 = redundant information + needs to be more explicit
- 410 = redundant information + sequence problem
- 411 = the question is not specific enough + format/layout problem + an explanation needs to be added + an example(s) needs to be added
- 412 = redundant information + explanation needs to be added
- 413 = title is not predominant + needs a summary at the end
- 414 = redundant information + syntax problem
- 415 = discrimination problem + typographic problem
- 416 = sentences are too long + sequencing problem + needs to be more explicit
- 417 = syntax problem + typographic error
- 418 = key concepts/terms need to be underlined + an advance organizer/preview needs to be added + format disrupts topicalization
- 419 = discrimination problem + example(s) needs to be more relevant
- 420 = needs to be more explicit + format/layout problem
- 421 = format/layout problem + the level of complexity needs to be increased

5. NON-CLASSIFIABLE RESPONSES FOR KINDS OF PROBLEM EXPLICATIONS

501 = content inaccuracy

502 = terminology is difficult to remember

503 = content inaccuracy + chunking problem + format/layout problem

504 = another medium is preferred

505 = it is difficult to learn and criticize at the same time

6. INABILITY OF SUBJECT TO EXPLICATE PROBLEM

601 = subject identifies (locates) but does not explain problem

Table 73 presents the modified Debert categorical system representing all of the revision responses, from all sources of feedback, which were generated from the study:

Table 73.

General Categories and Sub-categories for Revision Responses

1. COMPREHENSION/RECALL KINDS OF REVISIONS:

101 = add training on the use of the materials

102 = add an advance organizer/preview

103 = add illustration

104 = add an example(s)

105 = add a definition + add illustration

106 = simplify the language

107 = use smaller units

108 = rearrange the sequence

109 = remove irrelevant information

110 = make example more relevant

111 = relate title to content

112 = add a definition

113 = modify syntax

114 = reclassify some of the objectives

115 = add explanation

116 = add explanation + add example(s)

117 = elaborate + add example(s)

118 = clarify

119 = add an objective to the list

120 = be more specific + add examples

121 = be more specific

122 = use consistent definitions

123 = elaborate

- 124 = be explicit as to how many responses are required
- 125 = conditionalize item stem
- 126 = rearrange the sequence + relate title to content
- 127 = change response format of question
- 128 = reword item stem
- 129 = add a summary at the end
- 130 = add an example(s) + add illustration + add explanation
- 131 = add a glossary of technical terms + add illustration + rearrange the sequence
- 132 = delete an example
- 133 = add a glossary of technical terms

2. MOTIVATION/ATTENTION KINDS OF REVISIONS:

- 201 = add illustration
- 202 = remove irrelevant information
- 203 = increase level of complexity
- 204 = add test items
- 205 = make examples more relevant
- 206 = delete redundancy
- 207 = delete redundancy + increase level of complexity
- 208 = increase level of complexity + add test items
- 209 = relate content more to nursing
- 210 = add activities

3. PHYSICAL FEATURE KINDS OF REVISIONS:

- 301 = change typeface
- 302 = delete black lines
- 303 = use point form (list)
- 304 = highlight using capital letters
- 305 = underline important terms
- 306 = number the sections
- 307 = modify typographic format of the title
- 308 = correct typographic error
- 309 = highlight using capital letters + list in point format
- 310 = underline important words + modify topic heading
- 311 = modify topic heading
- 312 = label the diagrams
- 313 = number the sections + modify topic heading
- 314 = change the instructional design format
- 315 = widen the spacing between the 2 topics
- 316 = make other media available as well
- 317 = underline important terms + use point form (list)

4. MULTIPLE RESPONSES OVERLAPPING TWO OR MORE OF THE ABOVE KINDS OF REVISIONS:

- 401 = highlight using capital letters + add explanation
- 402 = delete redundancy + modify syntax

- 403 = delete black lines + add an advance organizer/preview
- 404 = correct typographic error + add explanation
- 405 = modify syntax + delete redundancy + add explanation
- 406 = delete black lines + make better transitions between sections
- 407 = delete redundancy + correct typographic error
- 408 = put in point form + add a definition + add an explanation
- 409 = delete redundancy + rearrange sequence
- 410 = reword item stem + change the instructional design format + add explanation + add examples
- 411 = modify typographic format of the title + add a summary at the end
- 412 = increase level of complexity + add a definition
- 413 = delete redundancy + use proportional spacing
- 414 = add examples + make examples more relevant
- 415 = add test items + add transfer exercises
- 416 = delete redundancy + add explanation
- 417 = add an example(s) + add illustration

5. NON-CLASSIFIABLE RESPONSES FOR KINDS OF REVISIONS:

- 501 = make content accurate
- 502 = change to celsius
- 503 = add another response to be considered as correct
- 504 = make content factual/accurate + change the instructional design format
- 505 = just read and do the test - then go back and criticize
- 506 = change the instructional medium

6. INABILITY OF SUBJECT TO MAKE/SUGGEST A REVISION

- 601 = a problem was identified but student was unable to make/suggest a revision

Tables 72 and 73 are the results of the content analysis of all of the comments that test subjects made during instruction and during debriefing. The six general categories for both problem explication and revision responses emerged from a close examination of the specific kinds of problem explications and revisions, i.e., the sub-categories as listed above. Similar kinds of problem explications and revisions were grouped together.

The modified Debert classification system presented in Tables 72 and 73 was used as input for the coding for the next two research questions to be discussed.

Research Question #6: Do experimental conditions affect the amounts of general kinds of problem explanations and revisions/revision prescriptions made during instruction?

For this research question all of the original verbatim problem explanation and revision/revision prescription responses (comments) which all of the test subjects made during instruction were classified using the modified Debert classification system (the outcome of having conducted the content analysis of all of the comments from all of the sources of feedback). For all of the 72 test subjects, each one of their original responses was placed (recoded) into one of the six general categories for both problem explanation and revision, e.g., if a subject's problem explanation response indicated that "an example needed to be added" then this statement would be placed into the general category of "Comprehension/Recall Kinds of Problem Explanation" Tables 72 and 73 presented these categories.

Frequencies of original verbatim responses placed into each general category for both problem explanation and revision were tabulated by general category for each test subject. The six general problem explanation and revision categories were conceptualized as dependent variables. As such, the SPSSX MANOVA procedure was applied to this tabulated frequency data, for each test subject, across the six general kinds (categories) of problem explanation responses and revisions/ revision prescription responses (the six dependent variables used in the analyses for each type of response). The two independent variables used in the analyses were the same as the previously discussed two-way ANOVA procedure: role of participants (with active and passive as the two levels of this independent variable) and number of participants (with one-to-one and small group as the two levels of this independent variable). Separate 2 x 2 multivariate analyses were conducted on both the amounts, across categories, of the general kinds of problem explanation responses and revisions/revision prescription responses. From this one can determine whether there was a significant difference among the tabulated amounts of the six kinds of categorized responses for both the general kinds of problem explanation responses and revisions/revision prescription responses. Separate 2 x 2 MANOVA's were conducted for Unit one responses, Unit two responses and their totals (Unit one + Unit two). Figure 8, which was presented in Chapter V, is a graphic representation of the data analysis for the statistical design. All descriptive

statistics related to the results of the 2 x 2 MANOVA of the data can be found in Appendix J. Summary tables and graphic portrayals of the results are incorporated into the text.

Figure 19 (see Appendix J) presents a histogram showing the frequencies of the amounts of the general kinds of categorized problem explication responses by unit for the total sample (N=72).

Figure 20 presents a histogram showing the frequencies of the amounts of the general kinds of categorized revision responses by unit for the total sample.

Tables 74 and 75 present the weighted means based on the 2 x 2 MANOVA of the effects of the independent variables on the amounts of the general kinds of categorized problem explication responses and the amounts of the general kinds of categorized revision/revision prescription responses for Unit one. Tables 76 and 77 are summary tables of the results of the multivariate analysis for Unit one. No significant differences were found for either the main effects of the independent variables or their interactions across either the amounts of the general kinds of categorized problem explication responses or the amounts of the general kinds of categorized revision/revision prescription responses.

Tables 78 and 79 present the weighted means based on the 2 x 2 MANOVA of the effects of the independent variables on the total amounts of the general kinds of categorized problem explication responses and the total amounts of the general kinds of categorized revision responses for Unit two. Tables 80 and 81 are summary tables of the results of the multivariate analysis for Unit two. Significant differences were found for the main effect of the independent variable of number of participants on the amounts of the general kinds of categorized problem explication responses, $F(6, 62) = 2.35, p < .05$. Significant differences were also found for the two-way interaction of the effects of the independent variables of number of participants and role of participants on the amounts of the general kinds of categorized problem explication responses, $F(6, 62) = 3.14, p < .01$. As Finn and Mattsson (1978) have noted, a significant multivariate main effect cannot be interpreted unless the results of the significant multivariate interaction present similar patterns of the effects of the independent variables across each one of the dependent variables. Figures 21a and 21b graphically portray the univariate interactions of the effects of the independent variables of role and number of participants on each dependent variable (the amounts of the general kinds of categorized

Table 76.

Summary of the 2 x 2 Manova Results for the Categorized Problem Explications for Unit #1

Source	df	Wilks F	P
Role of Participants (R)	6	1.55	0.18
Number of Participants (N)	6	1.24	0.30
R X N Interaction	6	0.70	0.65
Error	63		

Table 77.

Summary of the 2 x 2 Manova Results for the Categorized Revisions for Unit #1

Source	df	Wilks F	P
Role of Participants (R)	6	1.05	0.41
Number of Participants (N)	6	1.48	0.20
R X N Interaction	6	1.09	0.38
Error	63		

Table 80.

Summary of the 2 x 2 Manova Results for the Categorized Problem Explications for Unit #2

Source	df	Wilks F	P
Role of Participants (R)	6	1.28	0.28
Number of Participants (N)	6	2.35	0.04*
R X N Interaction	6	3.14	0.01**
Error	62		

* $p < .05$ ** $p < .01$

Table 81.

Summary of the 2 x 2 Manova Results for the Categorized Revisions for Unit #2

Source	df	Wilks F	P
Role of Participants (R)	6	1.96	0.09
Number of Participants (N)	6	1.49	0.20
R X N Interaction	6	1.77	0.12
Error	62		

Two-way Interactions

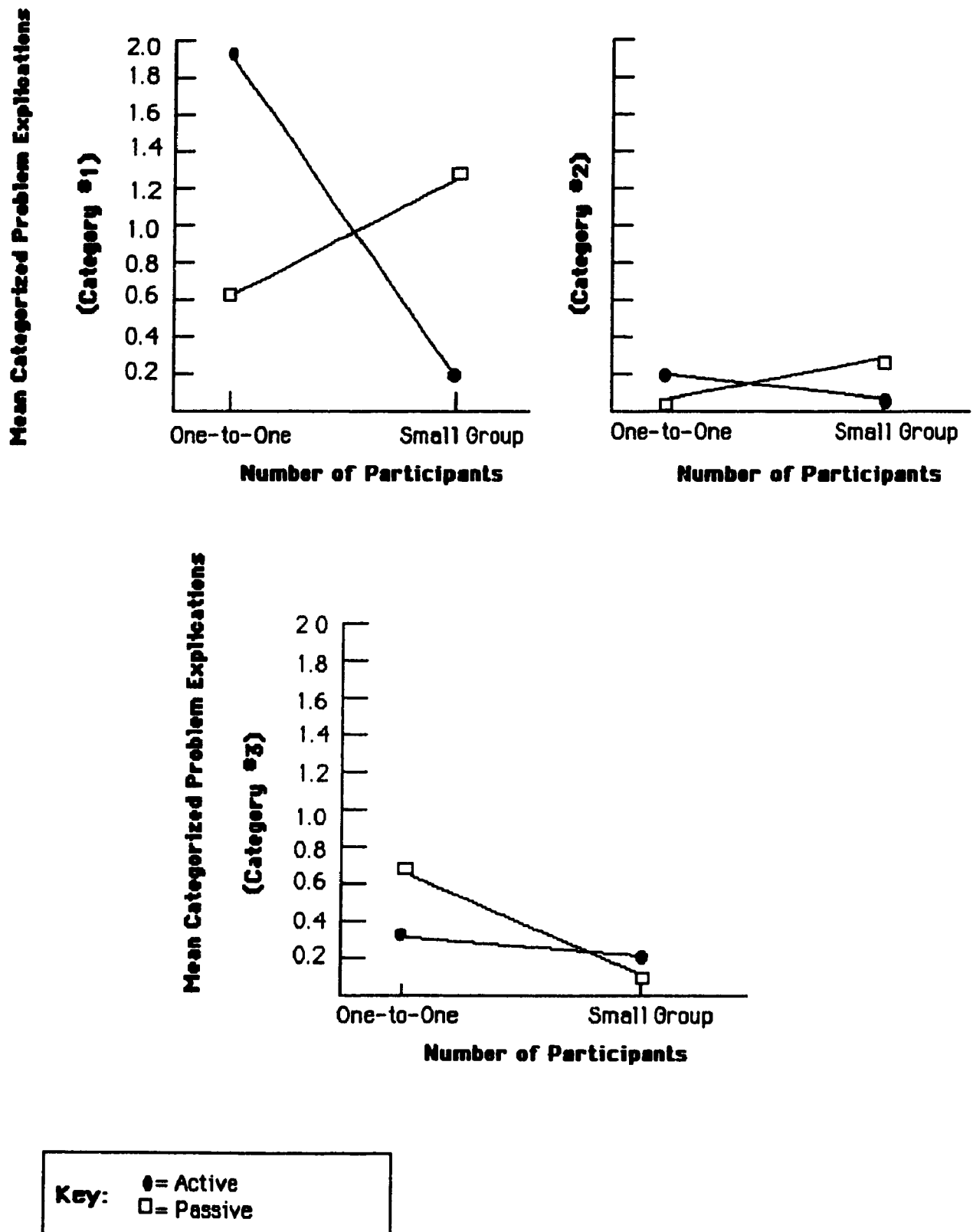


Figure 21 a. Plotting of means. Interactions of role and number of participants with categorized problem explanations for Unit #2.

Two-way Interactions

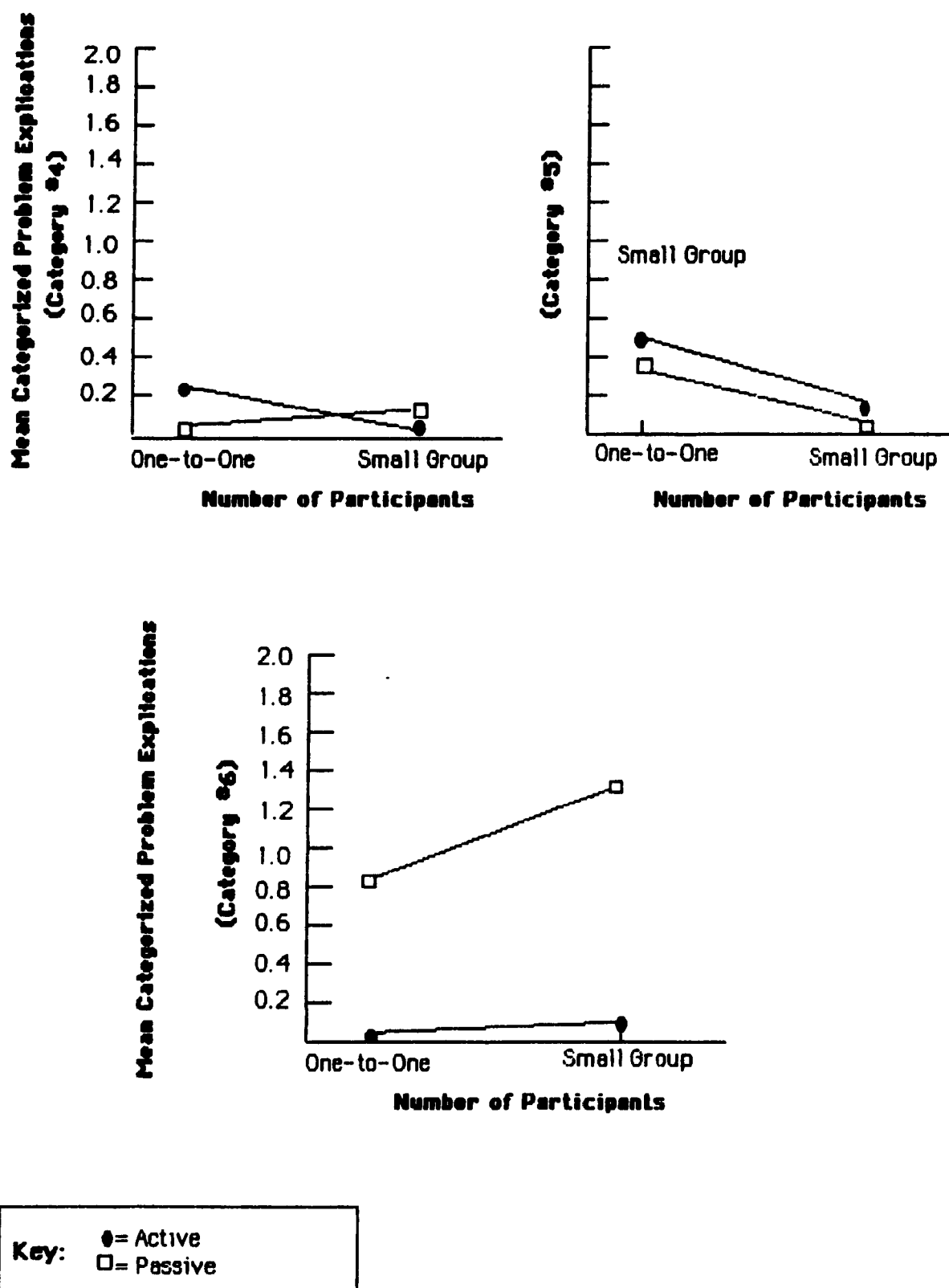


Figure 21 b. Plotting of means. Interactions of role and number of participants with categorized problem explanations for Unit #2.

problem explication responses). As Figures 21a and 21b demonstrate, the patterns of results are dissimilar for the levels of the independent variables across dependent variables for the significant multivariate two-way interaction of the effects of number of participants and role of participants; therefore the significant multivariate main effect of the independent variable of number of participants on the amounts of the general kinds of categorized problem explication responses can not be interpreted. With regards to the results of the univariate analyses, there was only a significant two-way univariate interaction of the effects of role and number of participants for amounts of general comprehension/recall kinds of problem explications (category #1), $F(1, 67) = 8.121, p < .01$. The results of the univariate analyses are summarized in Table 82. The one-to-one active condition produced more general comprehension/recall kinds of problem explications than any of the other experimental conditions. None of the other general kinds of problem explication categories had statistically significant results related to them.

Tables 83 and 84 present the weighted means based on the 2×2 MANOVA of the effects of the independent variables on the total amounts of the general kinds of categorized problem explication responses and the total amounts of the general kinds of categorized revision responses for Units one and two combined. The results of the multivariate analysis are summarized in Tables 85 and 86. No significant differences were found for the effects of the independent variables across either the amounts of the general kinds of categorized problem explication responses or across the amounts of the general kinds of categorized revision responses.

Research Question #7: Do experimental conditions in conjunction with the test subject characteristic of prior knowledge of the material affect the amounts of general kinds of problem explications and revisions/revision prescriptions made during instruction?

The final research question addressed by the study used the same MANOVA data analysis procedures reported earlier with the exception of the introduction of prior knowledge as a third independent variable with two levels (prior knowledge and no prior knowledge). The six general problem explication and revision categories again were conceptualized as dependent variables. As such, the SPSSX MANOVA procedure was applied to this tabulated frequency data, by experimental condition,

Table 82.

Two-Way ANOVA of the Amount by Category of General Kinds of Problems Explicated during the
Developmental Testing of Unit Two

Category: Source #1	Sum of squares	df	Mean squares	F	Significance of F
R x N Interaction	26.217	1	26.217	8.121	.006**
Error	216.281	67	2.228		

* p < .01

Note that:

R = Role of Participants

N = Number of Participants

Category: Source #2	Sum of squares	df	Mean squares	F	Significance of F
R x N Interaction	1.094	1	1.094	3.325	.073
Error	22.052	67	0.329		

Note that:

R = Role of Participants

N = Number of Participants

Category: Source #3	Sum of squares	df	Mean squares	F	Significance of F
R x N Interaction	0.705	1	0.705	1.465	.230
Error	32.248	67	0.481		

Note that:

R = Role of Participants

N = Number of Participants

Table 82 (continued).

Two-Way ANOVA of the Amount by Category of General Kinds of Problems Explicated during the
Developmental Testing of Unit Two

Category: Source #4	Sum of squares	df	Mean squares	F	Significance of F
R x N Interaction	0.334	1	0.334	2.278	.136
Error	9.830	67	0.147		

Note that:

R = Role of Participants

N = Number of Participants

Category: Source #5	Sum of squares	df	Mean squares	F	Significance of F
R x N Interaction	0.000	1	0.000	0.000	.984
Error	30.542	67	0.456		

Note that:

R = Role of Participants

N = Number of Participants

Category: Source #6	Sum of squares	df	Mean squares	F	Significance of F
R x N Interaction	0.850	1	0.850	0.268	.606
Error	212.542	67	3.172		

Note that:

R = Role of Participants

N = Number of Participants

Table 85.

Summary of the 2 x 2 Manova Results for the Total Number of Categorized Problem Explications

Source	df	Wilks F	P
Role of Participants (R)	6	2.15	0.06
Number of Participants (N)	6	1.55	0.18
R X N Interaction	6	1.90	0.10
Error	62		

Table 86.

Summary of the 2 x 2 Manova Results for the Categorized Revisions

Source	df	Wilks F	P
Role of Participants (R)	6	1.70	0.14
Number of Participants (N)	6	0.94	0.48
R X N Interaction	6	1.80	0.11
Error	63		

across the six general kinds (categories) of problem explication responses and revisions/revision prescription responses (the six dependent variables used in the analyses for each type of response). This was the same data that were analysed with the 2×2 MANOVA with the exception of having added the prior knowledge variable to the analysis. The three independent variables used in the analyses were: role of participants (with active and passive as the two levels of this independent variable); number of participants (with one-to-one and small group as the two levels of this independent variable); and prior knowledge (with prior knowledge and no prior knowledge as the two levels of this independent variable). Separate $2 \times 2 \times 2$ multivariate analyses were conducted on both the amounts, across categories, of the general kinds of problem explications and revisions/revision prescriptions (the same dependent variables reported for question six). Separate $2 \times 2 \times 2$ MANOVA's were conducted for Unit one responses, Unit two responses, and their totals (Unit one + Unit two). Figure 9, which was presented in Chapter V, is a graphic representation of the data analysis for the statistical design.

The 2×2 MANOVA data analysis design reported above for question six was processed separately from the $2 \times 2 \times 2$ MANOVA data analysis design (question seven). The 2×2 MANOVA data analysis design did have equal N's. The $2 \times 2 \times 2$ MANOVA data analysis design did not have equal N's due to the unequal distribution of test subjects along the dichotomized prior knowledge variable. This affected the computed values of the weighted means for the effects of the other two independent variables used in the $2 \times 2 \times 2$ MANOVA data analysis. The distribution of the unequal N's for the independent variables is indicated in Figure 9. All of the descriptive statistics related to the results of the $2 \times 2 \times 2$ MANOVA's can be found in Appendix K. Summary tables are incorporated into the text. The following results will focus on the prior knowledge variable and its relationship with the other two variables. Results relating to the main effects and interactions of the variables of role and number of participants are redundant with the outcomes of the 2×2 MANOVA and therefore will not be discussed.

Tables 87 and 88 (see Appendix K) present the weighted means based on the $2 \times 2 \times 2$ MANOVA of the effects of the independent variables on the amounts of the general kinds of categorized problem explication responses and the amounts of the general kinds of categorized revision responses for Unit one. Tables 89 and 90 are summary tables of the results of the multivariate analysis for Unit one. No

Table 89.

Summary of the 2 x 2 x 2 MANOVA Results for the Categorized Problem Explications for Unit #1

Source	df	Wilks F	P
Role of Participants (R)	6	1.08	0.38
Number of Participants (N)	6	1.08	0.39
Prior Knowledge (K)	6	1.05	0.40
R X N Interaction	6	0.66	0.68
R X K Interaction	6	0.37	0.90
N X K Interaction	6	0.90	0.50
R X N X K Interaction	6	0.73	0.63
Error	58		

Table 90.

Summary of the 2 x 2 x 2 MANOVA Results for the Categorized Revisions for Unit #1

Source	df	Wilks F	P
Role of Participants (R)	6	0.93	0.48
Number of Participants (N)	6	1.17	0.33
Prior Knowledge (K)	6	0.87	0.52
R X N Interaction	6	0.84	0.54
R X K Interaction	6	0.19	0.98
N X K Interaction	6	0.45	0.84
R X N X K Interaction	6	0.54	0.78
Error	58		

significant differences were found for either the main effects of the independent variables or their interactions across either the amounts of the general kinds of categorized problem explication responses or across the amounts of the general kinds of categorized revision responses.

Tables 91 and 92 present the weighted means based on the $2 \times 2 \times 2$ MANOVA of the effects of the independent variables on the amounts of the general kinds of categorized problem explication responses and the amounts of the general kinds of categorized revision responses for Unit two. Tables 93 and 94 are summary tables of the results of the multivariate analysis for Unit two. No significant differences were observed.

Tables 95 and 96 present the weighted means based on the $2 \times 2 \times 2$ MANOVA of the effects of the independent variables on the total amounts of the general kinds of categorized problem explication responses and the total amounts of the general kinds of categorized revision responses for Units one and two combined. The results of the multivariate analysis are summarized in Tables 97 and 98. No significant differences were found for either the main effects of the independent variables or their interactions for both the total amounts of the general kinds of categorized problem explication responses and for the total amounts of the general kinds of revision/revision prescription responses. Prior knowledge did not produce any significant main effects or interactions in the analyses reported above.

The next chapter will present a discussion of the above results by research question.

Table 93.

Summary of the 2 x 2 x 2 MANOVA Results for the Categorized Problem Explications for Unit #2

Source	df	Wilks F	P
Role of Participants (R)	6	1.05	0.40
Number of Participants (N)	6	1.98	0.08
Prior Knowledge (K)	6	1.11	0.37
R X N Interaction	6	2.62	0.03*
R X K Interaction	6	0.57	0.75
N X K Interaction	6	1.27	0.29
R X N X K Interaction	6	1.51	0.19
Error	58		

* $p < .05$

Table 94.

Summary of the 2 x 2 x 2 MANOVA Results for the Categorized Revisions for Unit #2

Source	df	Wilks F	P
Role of Participants (R)	6	1.82	0.11
Number of Participants (N)	6	1.29	0.28
Prior Knowledge (K)	6	1.71	0.14
R X N Interaction	6	1.81	0.11
R X K Interaction	6	0.64	0.69
N X K Interaction	6	1.16	0.34
R X N X K Interaction	6	1.85	0.78
Error	58		

Table 97.

Summary of the 2 x 2 x 2 MANOVA Results for the Total Number of Categorized Problem Explications

Source	df	Wilks F	P
Role of Participants (R)	6	1.62	0.16
Number of Participants (N)	6	1.30	0.27
Prior Knowledge (K)	6	0.71	0.64
R X N Interaction	6	1.86	0.10
R X K Interaction	6	0.44	0.85
N X K Interaction	6	1.01	0.43
R X N X K Interaction	6	1.35	0.25
Error	58		

Table 98.

Summary of the 2 x 2 x 2 MANOVA Results for the Total Number of Categorized Revisions

Source	df	Wilks F	P
Role of Participants (R)	6	1.34	0.26
Number of Participants (N)	6	0.82	0.56
Prior Knowledge (K)	6	0.96	0.46
R X N Interaction	6	1.71	0.13
R X K Interaction	6	0.47	0.83
N X K Interaction	6	0.73	0.63
R X N X K Interaction	6	0.91	0.50
Error	58		

CHAPTER VII

DISCUSSION OF THE RESULTS OF THE EXPERIMENT

A discussion of the results obtained and presented in Chapter VI will be presented for each research question in this chapter.

Research Question #1: Do experimental conditions affect the amount of data obtained from the different sources of feedback? and Do experimental conditions affect the amount of data obtained from the combined sources of feedback?

a) Comments During Instruction

It is interesting to note (Figure 10) that individuals assigned to small groups do not produce as many comments during instruction as the one-to-one procedure. This can be said consistently for Unit one, Unit two, and their combined totals. Figure 11 reveals that the active small group condition produced the fewest amount of comments overall. It is in this condition that test subjects were asked to declare the nature of their learning difficulties in front of a group of their peers. Possibly test subjects are reluctant to verbalize their problems in front of their fellow students for fear of being perceived as being "slow" or "less intelligent".

An examination of the descriptive statistics (Tables 11, 12, and 13) for the number of problems identified by the sample ($N=72$) during instruction for Unit one, Unit two, and the total for the two units combined reveals that a high percentage of the sample were non-responders. For Units one, two, and the two combined units respectively, 54.2%, 43.7%, and 36.1% of the sample did not identify any problems whatsoever. Means for the respective units show that 1.65, 2.47, and 4.08 problems were identified during instruction. The standard deviations for each of the units are all greater than their means. This demonstrates the extreme amount of variability among test subjects in identifying problems during instruction.

Contributing greatly to the mean values and the enormous variability, as indicated by the standard deviations, are individual test subjects who might be described as "high responders". Most of the test

subjects who did respond identified "a few" problems, e.g., one to five comments per unit. This contrasts sharply with high responding test subjects who identified "many" problems, e.g., 11 to 22 comments per unit.

Some test subjects appear to be either better able or more willing to identify their problems. If the intent of developmental testing is to collect as much feedback as possible, then instructional designers/evaluators should be aware of the extreme amount of variability of test subjects' ability and/or willingness to identify perceived problems within instructional materials. This finding is related to the four different experimental conditions employed in the study. Tables 20, 21, and 22 illustrate descriptively that within both of the one-to-one experimental conditions there is a high percentage of non-responders. Of those test subjects who are responders, most identify just a few problems. Then there are a few test subjects who identify larger numbers of problems.

In some ways small groups resemble one-to-one but they also differ in important ways. Like one-to-one, there is a high percentage of non-responders for both small group active and passive. Small group active produced an exceptionally high number of non-responders; 72.2% of test subjects assigned to this condition made no comments during instruction for both Unit one and Unit two. Small groups in general did not produce any high responders possibly due to intimidation.

It could be argued that high responders represent "ideal test subjects" in the sense that more data could be gathered from a sample of them than from a sample of non-responders. By selecting only high responders testing could be more cost effective. Fewer student test subjects would be needed to test the materials thus reducing the amount of time, money, and effort expended to evaluate instruction as it is being developed. Berthelot (1978) presents selection criteria for such highly verbalized test subjects.

The two-way ANOVA results demonstrate that one-to-one produced significantly more comments than small group for Unit one. Table 23 presents patterns of all of the results from the ANOVA's. The one-to-one condition consistently produced the greatest number of identified problems for Unit one, two, and for the combined total number of identified problems. Although the differences between the means for the second unit did not approach statistical significance, the mean difference between one-to-one and small group for the combined total number of identified problems closely approached statistical

significance, i.e., the probability of obtaining a mean value of 5.28 for one-to-one and a mean value of 2.89 for small group was computed to be .10. A perusal of the means for each one of the four experimental conditions reveals that within the one-to-one condition, one-to-one passive produced more comments than one-to-one active for Unit one. For Unit two and for the combined total, one-to-one active produced the greatest amount of feedback. Possibly one-to-one active subjects were inhibited to make comments during the first day with the experimenter, by their side, constantly observing and probing. One-to-one passive subjects had been left on their own. Perhaps the second day was less inhibiting for one-to-one active test subjects because by then they were familiar with both the experimenter and the experimental procedures. If the combined total is used as an overall index of the amount of identified problems (feedback), one-to-one active did produce the greatest number of comments of all of the experimental conditions.

b) Debriefed Comments

As with comments during instruction, debriefing comments showed tremendous variability. In part, this similarity is related to how the debriefing agendas were constructed. The problems that were identified during instruction were used as input for the construction of the debriefing agenda. This procedure obviously parallels the procedure used to gather comment data during instruction and might produce parallel results for the two sources of feedback data; however there were different criteria used to both construct the debriefing agendas and to record the debriefing outcomes between the one-to-one and small group conditions. In addition test subjects were allowed to identify new problems during debriefing that were not perceived during instruction. For small groups, a new problem would only be discussed and recorded as an outcome if at least two or more members in a group agreed that the new problem was indeed problematic.

Using Table 37 descriptively, the one-to-one active condition produced the greatest amount of debriefing outcomes in total. Small groups may have been inhibited to respond during debriefings because the experimenter could not follow all lines of inquiry at once. By the experimenter probing further along one line another line of inquiry might be excluded. The lack of responding among a group

may not indicate that everyone is agreed. A group member may just drop his or her comment because the discussion has passed that point.

c) Debriefed Self-Test Errors

Significantly more self-test items, based on errors made, are debriefed with the small group procedure than for one-to-one for Unit one and Unit two and in total. In reference to Figure 14, which plots the means of the main effects of number of participants, this result is surprising in that it was consistently more difficult for a self-test item to appear on a small group debriefing agenda for each unit. At least two of the six test subjects in a small group had to have gotten the item wrong in order for that item to appear on the agenda to be discussed. No such criterion was applied to one-to-one test subjects' errors and in this sense it is remarkable that so many more self-test items were debriefed in the small group condition. It appears that if the input to revision includes both comments and error data, more comments will be gained from the one-to-one condition but more error data will be obtained from small groups.

d) Attitude Survey

Frequency data presented in Tables 62, 63, and 64 illustrate that virtually all conditions produced critical feedback based on debriefing the results of the attitudinal survey. Perhaps the relative anonymity of filling out a survey reduces a perception of intimidation within a small group to identify one's learning problems. This result contrasts sharply with the results of the analysis of the second source of feedback: debriefed problems (comments). Small groups were generally reluctant to identify their problems. In a small group, the attitudinal survey provided a relatively innocuous way to indicate perceived problems when compared to being requested to either write out or verbalize one's problems in front of peers.

Poor self-test results and poor attitude results suggest that small groups were having problems not revealed by comments during instruction.

Do experimental conditions affect the amount of data obtained from the combined sources of feedback?

Why were no significant differences between experimental conditions detected for the total number of debriefing outcomes (comments + self-test errors + survey responses) over the two days especially when significant differences were found for each of the separate analyses of these sources of feedback data? This result is explained by looking at the results of each analysis by source of feedback. For debriefed comments one-to-one produced a much greater amount of feedback data than small group. For debriefed self-test results just the opposite was true, i.e., small group produced a much greater amount of feedback data. For debriefed attitudinal survey results, small group produced more feedback data than one-to-one although the differences are not as striking as was observed for the other two sources. The reason that the two-way ANOVA did not identify any significant differences is due to the variable contributions of experimental conditions, by source of feedback, to the amounts of feedback data. When all of the amounts of feedback data from the three sources of feedback are pooled previous effects by individual source are nullified or in other words previous effects between sources end up cancelling each other within the two-way ANOVA of the pooled data. It is for this reason that this result should be interpreted with caution — previous differences by each source were observed and best describe the results of this aspect of the study. Table 70 illustrates for descriptive purposes that the small group passive condition did produce the greatest amount of overall feedback data with the least amount of variability between passive small groups as revealed by the standard deviation. All of the other three experimental conditions produce less overall feedback and all have extremely large standard deviations.

Research Question #2: Does the test subject characteristic of having indicated prior knowledge of the material affect the overall amount of comments being made during instruction?

The results of the t-tests suggest that more comment data can be obtained from test subjects who possess the characteristic of having acquired prior knowledge of the material being tested. Of course, these test subjects cannot be considered to be representatives of the general target population because they are familiar with the material. Possibly prior knowledge allows these test subjects to be more critical of the materials. Less attention might be paid to learning while more attention could be spent

on assuming the role of critic. It should be noted that self-test and/or posttest results (error data) from a prior knowledge group would not be as useful as the results from representative test subjects because, presumably, few errors would be made. This would result in the loss of valuable feedback based on test results. Achievement data is best gotten from novices. Critical comments could be obtained from novices and those with prior knowledge. The primary goal of developmental testing is to empirically test the materials on the target population for whom the materials were designed. If members of the target population are not used during developmental testing then the instructional designer/evaluator will never know the effects of the instruction on the group for whom the instruction is targeted. A possible compromise might be to revise based on the feedback from test subjects with prior knowledge and then test out the revised instruction on the target group.

It is important to note that even though the amounts of feedback from the various sources have been quantified and analyzed that more feedback might not necessarily mean better feedback. This also applies to the findings related to the t-Test results between prior knowledge and no prior knowledge groups where, generally, the prior knowledge group produced more comments. The only way in which to determine the effectiveness of the different amounts of feedback would have been to have actually revised the materials based on the feedback from the different experimental conditions and to experimentally administer both the revised versions of the material and the unrevised version (using a control group). Attitudinal survey results and self-test results could be used as indicators of effectiveness. The amount of comments (problems) should also decrease. These are possibilities which go beyond the scope of the research being presented.

It should be reiterated that members of the potential target population who have mastered the objectives are not representative of that population. In speculation, test subjects with prior knowledge might supplement data gathered from student test subjects who are representative of the target population. If the developmental testing of instruction by learners is defined as an empirical approach to obtain feedback as to the effectiveness of the materials while the material are still considered to be prototypic then members of the target population must be used. However it seems reasonable to use, for example, graduated students of the program in which the instructional materials are to be used.

Graduated students who are in the work force and using the knowledge and skills gained from the educational program might be able to offer useful and current suggestions about instructional materials. These "quasi-expert" test subjects might be able to use the experience they have gained on the job in order to suggest ways in which the materials might be improved.

Research Question #3 To what extent do data from the various sources of feedback provide redundant or unique information? This is a general question addressed by more specific questions:

Specifically: a) *After instruction, as a debriefing strategy, to what extent does the procedure of having test subjects go over their comments about the materials again (in retrospect) produce unique or redundant feedback data?*

b) *To what extent does performance on individual self-test items reflect comments (perceived problems) made during instruction?*

c) *What are test subjects' perceptions of self-test error causation?*

d) *After instruction, as a debriefing strategy, to what extent does the procedure of having test subjects review their self-test errors produce unique or redundant feedback data?*

e) *After instruction, as a debriefing strategy, to what extent does the procedure of having test subjects review their unfavorable attitudinal survey responses produce unique or redundant feedback data?*

a) Unique or Redundant Feedback Before and During Debriefing

It can be concluded that the debriefing of comments made during instruction provided a considerable amount of novel feedback.

Although no data were analyzed on this point, informal observations suggest that small group debriefings contributed extensively to providing unique data. During the small group discussions, test subjects who had not indicated having perceived the problem in question were often persuaded, by those who did, that there was a problem. Sometimes even those who had not originally perceived the problem would offer problem explications and revisions that the group would agree to. Sometimes the test subjects who had originally perceived the problem and provided problem explications and revisions of their own would favour a completely different problem explication and revision suggest by someone who had not indicated the problem. Even for one-to-ones, a problem that was not perceived during instruction would be brought up during debriefing. Sometimes a problem perceived during instruction

was later resolved during instruction. Even if the results of debriefing had all been redundant with the comments made during instruction, it could be argued that confirmatory feedback data would be useful to collect. It can be concluded that the results have indicated that more than half of the feedback gathered by having test subjects review their comments is unique. Debriefing provides a useful mechanism to finalize the results of developmental testing for a test subject or group of test subjects.

b) Unique or Redundant Feedback: Self-Test Results and Comments During Instruction

The results of tabulating the frequencies and percentages presented in Chapter VI for the category of a problem being identified with the instruction but the related self-test item was responded to correctly and the category of no problem having been identified with the instruction and the related item was responded to incorrectly are of special interest. Responses falling into these two categories raise the issue of whether or not perceived problems noted during instruction can be used to verify performance on the self-test.

It appears to be a reasonable assumption that a problem encountered during instruction might be highly related to poor performance on a self-test item. Frequency data for the first category, relating an identified problem with the instruction to a correct response on the corresponding self-test item, might be explained by the test subject merely overcoming the problem. The clarification process which test subjects went through in order to explicate problems might have aided in order to overcome some of the problems. Frequency data for the third category, relating failure on the self-test item with failure to identify a problem with the corresponding instruction, can possibly be explained by poor alignment of the item with the instructional material. Or, perhaps students are poor detectors of problems with instruction. Regardless of the possible reason(s), the frequency data for category three reveal that test error data greatly contribute to the amount of input data during developmental testing.

c) Subjects' Perceptions of Causation of Self-Test Errors

The results of this question also indicate that test subjects are more likely to attribute errors made on the self-test to the instructional text rather than the actual self-test items. Perhaps test subjects are reluctant to criticize test items and provide problem explications and revisions because they have never been given this opportunity before and/or they feel they do not have the technical expertise to do so. It might be argued that in order to revise a test item that a critic should be well versed in the principles of test construction. Of the total number of 389 debriefed self-test errors, test subjects most frequently reported (47.56%) that there were no problems either with the instructional text or the self-test even though errors were made. Typically test subjects would state anecdotally that "they forgot" thus attributing the error to themselves rather than to the test item and/or the instructional material.

d) Unique or Redundant Feedback: Debriefing Comments on Problems Noted on the Self-Test

Based on the presented results it can be noted that having test subjects review their self-test errors provides a great deal of novel feedback. Test subjects were unable to identify problems while writing the test — they merely attended to the task of responding to items. In this way they were unable to assume the role of a test critic and merely behaved as traditional students. Of the 389 debriefed self-test errors only seven were redundant with problem explications and six for revisions made while actually writing the test.

e) Review of Comments on Survey Responses

The results of both the analysis of redundant problem explication and revision debriefing outcomes based on the results of the attitudinal survey are similar. It can be concluded that substantial amounts of unique feedback data were obtained by having the test subjects go over their unfavorable attitudinal responses for purposes of obtaining elaborations as to why they responded unfavorably.

It was common for a test subject or group of test subjects to have indicated dissatisfaction on the attitudinal survey but when it came time to discuss during debriefing why dissatisfaction had been indicated, test subjects changed their minds. What had been perceived as a problem while filling out

the survey was no longer perceived as a problem during the debriefing. Of the total of 813 unfavourable responses to the survey, 245 (30.1%) had indicated that what was a perceived problem while filling out the survey was no longer a problem during the debriefing discussion. The results, naturally, are very similar for revision responses. Based on these results alone, it can be concluded that the usefulness of having the test subjects go over their unfavorable responses has been empirically demonstrated. If an instructional designer/evaluator does discuss the results of the attitudinal survey with the test subjects, valuable and novel feedback can be gained and unwarranted problems and revisions can be avoided.

Research Question #4: Do subjects just locate and not explain problems within the instructional materials or can they explain and/or remediate the problems?

With regards to the 298 verbatim comments which represent all of the debriefing outcomes based on the problems identified during instruction, 39.6% of these student-generated comments involved the location, explication, and resolution of a perceived problem — certainly a sophisticated kind of a response. A substantial proportion of student-generated comments (18.46%) did not address problem explication but completely overrode the problem explication task and immediately generated a revision/revision prescription — also a sophisticated kind of a response. By pooling the frequencies of both comments involving the location, explication, and resolution of a perceived problem and comments which only generated a revision/revision prescription it can be shown that 173/298 or 58% of the debriefing outcomes represented a revision/revision prescription. It can therefore be concluded that most students can do more than just physically locate possible problems and in fact most can suggest a way to revise the instruction. It should also be noted that 16.78% of student comments were at the level of just locating a perceived problem without being able to explicate and/or remediate the problem — certainly a much less sophisticated kind of response. However, from an instructional designer's point of view, even having this kind of feedback is more desirable than not being aware that there might be a problem. Finally, it is noted that new problems that were undetected during instruction were identified during debriefing. Although only 5/298 or 1.68% of the total number of debriefing outcomes based on

comment data were unique to the debriefings, it is recommended that new comments be attended to during debriefings.

Research Question #5: What general and specific kinds of problem explications and revisions/ revision prescriptions result from developmental testing and debriefing?

As Tables 72 and 73 (presented in Chapter VI) reveal, most of the specific problem explications and revisions (the sub-categories) are related to comprehension and recall of the instruction. On some occasions, test subjects would generate a multiple response overlapping two or more of the three main general categories of: 1. Comprehension/Recall; 2. Motivation/ Attention; and 3. Physical Features. Some comments were placed into the Non-Classifiable Responses general category. As an example, sometimes test subjects would dispute the truth value of propositions within the material.

Examination of all of the specific kinds of problem explication responses, i.e., the sub-categories in Table 72, reveals that student test subjects are able to comment about a wide variety of problems they perceive within the materials.

Student test subjects generally tended to approach the problem explication task in two different ways. One kind of approach generated a response more along the lines of a solution-oriented problem explication, e.g., "needs more illustration". The other kind of approach was more of a straight forward problem statement, e.g., "I don't understand".

Examination of the sub-categories in Table 73, similarly reveals that student test subjects are able to suggest a wide variety of solutions to the problems they perceive within the materials. When asked to suggest a way of avoiding the problem so that other students would not encounter the same problem, students generally tended to respond by either suggesting a way to remediate the problem, i.e., prescribe a revision, or by actually rewriting/revising the instruction.

The category of problem explication did not dictate the category of revision. For example, when test subjects indicated that they did not understand something, a problem explication placed into the comprehension recall category, the complementing proposed revision might have been placed into the physical features category. For this reason, no attempt should be made to directly relate the sub-categories for problem explication to the subcategories for revision. It should also be noted that there

are instances when there are overlapping entries for some of the sub-categories. For example, "needs more illustration" appears in both the general category of comprehension/recall and the motivation/attention category for problem explication. Based on the transcripts of the sessions, the discrimination was made during the content analysis as to which category the response belonged in based on the subject's explication of the problem.

Several general conclusions can be drawn based on the results of the content analysis. Student test subjects are able to both explicate perceived problems and suggest ways of resolving these perceived problems. The sub-categories attest to this. They also are examples of kinds of problems that an instructional designer/evaluator would attend to during the developmental testing of prototype self-instructional materials in order to improve the materials.

It might be useful, for training learners to be test subjects, to provide them with a list of examples of the kinds of problem explications/revisions that tend to be made based on developmental testing. Debert's list of revision prescriptions or the modified versions based on his work might be useful. This training aid could be used to complement general directions given to test subjects in order to insure that they comprehend the tasks involved in being a test subject.

Research Question #6: Do experimental conditions affect the amounts of general kinds of problem explications and revisions/revision prescriptions made during instruction?

Although the results of the analyses detected no significant differences for Unit one and for the totals of the two combined units, significant differences were found for Unit two. The one-to-one active condition produced more comprehension/recall kinds of problem explications than any of the other experimental conditions. Possibly this result is merely serendipitous and unstable. If the effect were stable then similar results would have been found also for Unit one and for the total of the two combined units. With regard to the amounts of feedback data for the first category of problem explication, a perusal of the tables of means for Unit one and for the total of the combined units (Tables 74 and 83) reveals that the means for the one-to-one active condition are consistently greater than any of the other experimental conditions for the amounts of comprehension/recall kinds of problem explications. This does lend some support for the conclusion that the one-to-one active condition generally produces more

of this kind of feedback. Experimental conditions affected only the amount of comprehension/recall kinds of problem explications. None of the other categories of problem explications had significant results related to them.

It might be argued that comprehension recall kinds of problems are the most important kinds of problems to be detected during developmental testing. If learners do not understand the subject matter and can not recall important facts or concepts then it is likely that the instruction has failed for these learners. Learners will certainly do poorly on the self-test if they did not understand the material and can not recall important aspects of the material especially on self-tests that largely measure recall performance (as the two self-tests employed in the experiment did). The one-to-one active condition consistently identified more of these extremely important kinds of problems during instruction. As suggested earlier, probably this condition was the least intimidating of all of the experimental conditions. Also test subjects were probed on a page-by-page basis. This probe was intended to remind the test subjects in this condition that the experimenter welcomed any kind of critical feedback which the subject might wish to express. Test subjects were also probed whenever the experimenter had perceived that a test subject might be experiencing some kind of difficulty. This again reinforced to the test subjects that the experimenter was very concerned about identifying any kind of a problem being encountered while the test subject was trying to learn the material. The test subjects in the one-to-one active condition often saw the experimenter carefully write down on the materials exactly what the test subject was articulating as to the nature of the problem. Possibly for these reasons, the one-to-one active condition facilitated the collection of more student feedback than any of the other experimental conditions.

Although small group results had no significant results related to them, small groups should not be ignored. It is interesting to compare small group passive with small group active. A perusal of Tables 74, 75, 78, 79, 83, and 84 consistently show suppressed responses for test subjects assigned to the small group active condition. Small group passive test subjects generally provided more kinds of feedback data than small group active subjects but by no means as much as one-to-one active. Perhaps being placed in a situation where a subject has to verbally declare their problems in front of their

peers is more intimidating than just having to write the problems down on the materials. In addition, there is another potentially inhibiting effect operating during the small group active condition. The experimenter was only able to attend to one test subject's problem at a time. Perhaps another group member experiencing problems while the experimenter was with a test subject merely went on with the materials rather than wait until the experimenter was free.

Descriptive statistics presented in Figures 19 and 20 greatly help in gaining an understanding of what test subjects can be expected to produce when they are presented with the tasks of both explaining and resolving their perceived problems during instruction. Figure 19 shows the frequencies of responses falling into each of the six general kinds/categories of problem explications. Test subjects in general identified more comprehension/recall kinds of problems than any other kind for Unit one, Unit two, and their totals. The second most frequently occurring kind of problem explication was for the sixth category. A substantial proportion of the responses made by test subjects indicated that they were not able to explain their problems during instruction after they had physically located the problem within the materials. Anyone conducting developmental testing should be aware that a test subject's ability to spot a problem is not necessarily complemented with the ability to explain the sources of the problem.

Figure 20 shows the frequencies of responses falling into each of the six general kinds/categories of revisions. No significant differences between experimental conditions were detected for categorized revision response data. If the experimental conditions are disregarded, then Figure 20 reveals that test subjects in general produced more revision/revision prescriptions related to the physical features of the materials than any other kind. This can be said for Unit one, Unit two, and their totals. Category six has the highest frequencies, although responses falling into this category did not produce a revision; instead test subjects just indicated that they could not resolve the problem when requested to. From this it can be concluded that there will be times during developmental testing when a test subject will just not be able to make or suggest a revision. This should be acknowledged by anyone conducting developmental testing. Test subjects eliciting this kind of a response should be asked to move on within the materials. When test subjects are able to make or suggest a revision it is likely to be related to the

more superficial aspects of physical features and format of the material. As a comparison between the problem explication frequency data, by category, and the revision data it is interesting to note that although test subjects most frequently identify comprehension/recall problems they most frequently tend to suggest a revision related to the physical features of the instructional material. It is important to be aware of the limitations of test subjects' ability to actually revise instruction. The above results point to the tendency of test subjects to deal with the more superficial aspects of the text. A more appropriate revision might have to be at a deeper structural level in order to resolve encountered problems.

Based on the results of all of the above multivariate analyses, significant differences between the experimental conditions were found only between the amounts of the general kinds of problem explications and not revisions/revision prescriptions. The frequency distributions presented in Figure 20 reveals that many responses indicated that the 72 test subjects had difficulty in resolving their identified problems. The inferential statistical tests confirm that this applies equally to all experimental conditions.

Research Question #7: Do experimental conditions in conjunction with the test subject characteristic of prior knowledge of the material affect the amounts of general kinds of problem explications and revisions/revision prescriptions made during instruction?

The primary interest in computing the $2 \times 2 \times 2$ MANOVA was to determine the main effects and interactions of the independent variables associated with the prior knowledge variable. It can be concluded that no significant results were observed for any of the main effects or interactions related to this variable. The previous results reported in this study concluded that test subjects who had indicated prior knowledge did produce significantly more comment data during instruction than test subjects who did not indicate prior knowledge (based on the results of the t -tests). The results of the MANOVA have indicated that when the feedback data is broken down by kind test subjects who had indicated prior knowledge did not produce significantly more feedback than test subjects who did not indicate prior knowledge. No particular kind of comment or revision stood out, in terms of amount for the prior

knowledge variable. This result is not related to results found by the t-test. The total amount of comment data analyzed in the t-test was subdivided into six categories and it was with the amounts of these kinds of data that the MANOVA was applied.

Some General Methodological Problems

Usually the inferential statistic of ANOVA/MANOVA is associated with hypothesis testing. In this study there were no research hypotheses — there were descriptive research questions guiding the investigation. There are also methodological problems in interpreting the above results and conclusions based on the experimental design. It is difficult to compare the results of Unit one and Unit two. They could be conceived of as two different experiments. The results of Units one and two are contaminated by a number of uncontrolled factors. Although the experimental procedures were the same for both units, Unit two was a much longer lesson than the first. The length of the second self-test reflects this as well as the differing amounts of time it took to go through the units. Differences in response frequencies by test subjects across the two units could be attributable, for example, to there being inherently more problems in one unit than the other — this is not known. Or, there might be a group dynamics effect or peer effect between the two units (tested over two days) because subjects in small groups become less inhibited in making responses after becoming more familiar with each other and the experimenter. The empirical (c.f., McGrath & Altman, 1966) and theoretical literature (c.f., Corey & Corey, 1982) on group processes suggests there are phases through which small groups are thought to develop. Possibly subjects are more aware of what is expected of them (a training effect) during developmental testing for the second day. One can only speculate on possible contaminating factors

Some of the results of this study might be construed as having implications for some broad questions about developmental testing and revision. For example, when should a developer/evaluator cease developmental testing? Rather than specify a specific number of one-to-one and/or small groups based on empirical results, the question might be addressed from a more theoretical position. Perhaps developmental testing should only stop after test subjects (in either one-to-one or small group

conditions) meet the cognitive, attitudinal, and/or psychomotor objectives for which the instruction was designed. It might be recalled that based on the results of the survey of expert practitioners' procedures for testing and revising that there were no significant differences between those experts using small group testing exclusively and those experts using both one-to-one and small group methods. If the instruction is truly designed to be used on a self-instructional basis, then the one-to-one phase of formative evaluation can not be eliminated because this manner of testing is similar to the way in which the materials were intended to be used. It has also been concluded, based on the results from the experiment that different kinds of data are obtained during instruction from using both of these testing methods or strategies. Anecdotally speaking, the interaction between group members during the debriefing sessions was much like a group problem-solving task or brain storming activity. This resulted in the generation of a large number of creative suggestions for both explaining and remediating perceived problems. Even subjects who had not perceived the problem during instruction, at times became convinced that there were problems based on other group members' problem explications and revisions. It is this kind and amount of interaction that can not be gained from one-to-one testing.

The next chapter will present the conclusions of the experiment and the survey of expert practices for designing and conducting developmental testing sessions. Conclusions from the results of the experiment will be related to the results of the findings of the survey.

CHAPTER VIII

CONCLUSIONS

This chapter will present the conclusions of both the experiment and the survey of expert practices for designing and conducting developmental testing sessions. Where appropriate, conclusions from the results of the experiment will be related to the results of the findings of the survey.

In order to integrate and summarize the results and conclusions of both the survey and the experiment, each of the research questions posed for the experiment will be presented and responded to based on all of the results which have been presented.

Do experimental conditions affect the amount of data obtained from the different sources of feedback?

a) Comments During Instruction

Some of the most interesting and important findings of this study are revealed through the descriptive statistics. There was an extremely large number of test subjects in all conditions who provided no comments during instruction. With regard to frequencies for the number of identified problems for the total sample, for Unit one 54.2% of the total sample made no comments whatsoever during instruction; for Unit two, 43.7% of the total sample made no comments; and in total for the two units combined, 36% of the sample provided no feedback during instruction. The whole purpose for designing and conducting developmental testing sessions is collecting such feedback data.

The small group active contributed the least amount of comments during instruction.

The one-to-one active condition generally produced the greatest amount of feedback with respect to generating comments during instruction. These results appear to confirm the results from the data analysis of the survey where a large majority of the respondents indicated that active roles were

assigned to the participants in order to facilitate interaction, both during instruction and debriefing, between the test subject and the designer/evaluator.

b) Debriefed Comments

The one-to-one active condition produced more debriefing outcomes based on the comments made during instruction. If small groups were to be used either exclusively or in combination with one-to-one sessions, it can be concluded that imposing criteria on comments elicited during instruction for purposes of either setting an agenda or recording debriefing outcomes will be too restrictive. Small groups produced the least amount of comment data during instruction. For this reason all test subjects' comments in a group should be attended to during the debriefing. Otherwise, potentially valuable comments will be suppressed.

Debriefing was conducted by a large majority of the respondents to the survey.

c) Debriefed Self-Test Errors

In general, the criterion for setting the small group debriefing agendas, based on self-test errors, does not appear to be as constrictive as it was for comments. Significantly more self-test items were debriefed using the small group procedure than the one-to-one procedure even though criteria imposed on small group self-test results made it more difficult for an item to be placed on the agenda. Results of the survey of expert practices indicated that criterion test results were considered to be an important source of feedback and some experts even rely exclusively on test results. The small group procedure might inflate the amount of debriefing outcomes based on self-test results due to the testing environment. Self-instructional materials are designed to function auto-instructionally but during small group testing test subjects are using the materials in a setting that does not represent the manner in which the materials were intended to be used. This might have contributed to the extremely high number of small group debriefing outcomes related to the self-test results.

d) Attitude Survey

It can be said that the attitudinal survey produced considerable amounts of feedback for all conditions. Actually, as a source of feedback, the attitudinal survey produced more feedback than any other source, i.e., outcomes based on the comments or the self-test results. This is probably the most interesting finding from the experiment and contrasts sharply with the procedures to which the expert practitioners adhere. The expert instructional designers indicated that they did not formally administer an attitudinal survey. As revealed by the results of the experiment, instructional designers will lose a large amount of potentially useful feedback if they do not utilize the results from this kind of an instrument. It appears that due to the intimidating nature of placing test subjects into the novel role of an instructional critic that many might be reluctant to reveal the problems they are encountering during learning. This especially applies to test subjects placed in small groups where it is difficult to provide anonymity or confidentiality for this kind of feedback. The attitudinal survey does provide this kind of anonymity and might be useful for test subjects who are reluctant to verbalize their perceived problems. The experiment revealed that the survey was an extremely important source for obtaining feedback data — in fact the most important of all of the sources in terms of the overall amount of feedback obtained.

For the number of debriefing outcomes based on the attitudinal survey results, the small group condition produces more feedback data than the one-to-one — especially the small group passive condition.

Does the test subject characteristic of prior knowledge of the material affect the overall amount of comments being made during instruction?

Possibly prior knowledge allows test subjects to be more critical of the materials. The t-tests between the groups with and without prior knowledge did indicate that the subjects with prior knowledge commented more frequently. Feedback from test subjects who are familiar with the material can be used to supplement data from members of the target population.

To what extent do data from the various sources of feedback provide redundant or unique information?

This is a general question addressed by more specific questions:

- Specifically:
- a) After instruction, as a debriefing strategy, to what extent does the procedure of having test subjects go over their comments about the materials again (in retrospect) produce unique or redundant feedback data?
 - b) To what extent does performance on individual self-test items reflect comments (perceived problems) made during instruction?
 - c) What are test subjects' perceptions of self-test error causation?
 - d) After instruction, as a debriefing strategy, to what extent does the procedure of having test subjects review their self-test errors produce unique or redundant feedback data?
 - e) After instruction, as a debriefing strategy, to what extent does the procedure of having test subjects review their unfavorable attitudinal survey responses produce unique or redundant feedback data?

As there are many sources of student data, a cost effective approach to developmental testing would determine the potential addition each could make and utilize the most effective mixture.

The results of the experiment revealed that having students review the comments they made during instruction afterwards, in a debriefing setting, produced substantial amounts of new feedback. For problem explications resulting from the debriefing of comment data, 51.4% were non-redundant/unverified with the original input for debriefing while 55.5% of the revision responses were non-redundant.

The results of the experiment revealed that having students review their self-test errors during debriefing produced substantial amounts of new feedback. For problem explications resulting from the debriefing of self-test errors, 98.2% were non-redundant/unverified with the original input for debriefing while 98.5% of the revision responses were non-redundant.

The results of the experiment revealed that having the test subjects review, during debriefing, unfavorable responses to an attitudinal survey produced substantial amounts of new feedback. For problem explications resulting from the debriefing of unfavorable survey responses, 24.23% were

non-redundant/unverified with the original input for debriefing while 25.95% of the revision responses were non-redundant.

Clearly, debriefing, based on all of these sources of feedback contributed substantially in gathering the feedback data.

It might be recalled that the experts who had responded to the survey of developmental testing procedures indicated that debriefing provided the best data for purposes of revising the material. It can be concluded, from the results of this study, that debriefing certainly provides the most amount of data. Debriefing can be conceptualized as a mechanism by which the results of developmental testing are finalized for a given test subject or group of test subjects. It is worth recalling that having debriefed the unfavorable responses to the attitudinal survey resulted in test subjects changing their opinions. Of the 813 unfavorable responses which were discussed, 245 (30.1%) debriefed responses resulted in test subjects changing their minds, i.e., what was perceived as a problem while filling out the survey was no longer a problem. This kind of feedback is obviously useful when it comes time for revisions to be implemented. It is also worth noting that even for subjects who made no comments during instruction, that feedback from these subjects was obtained through the administration of the attitudinal survey. As an anecdote, the experimenter found that unfavorable responses to a survey item provided a convenient departure point, for debriefing purposes, to focus on the general kinds of perceived problems.

Do subjects just locate and not explain problems within the instructional materials or can they explain and/or remediate the problems?

Test subjects certainly can do more, when requested, than just pinpoint where the problem might be located. Most of those test subjects who responded could locate and explain their problems. Going beyond this, most of these test subjects suggested a way of overcoming the problem. It appears that test subjects are able to provide the kind of data recommended by the instructional design literature. Results of the survey of practitioners indicated that test subject identification of problems was their most important role during developmental testing. What conflicts with survey results is test subjects

suggesting or making revisions to the material. Very few respondents to the survey placed test subjects into an active revision role. The results of the experiment indicate that test subjects are able to suggest/make revisions however most test subjects will suggest superficial changes. The most frequent kinds of revisions made by test subjects during the experiment were related to the physical features of the instructional material. A few test subjects were able to go beyond the superficial aspects. It can be concluded that until a way to identify proficient student revisers is found that students assume a passive role for revision.

What general and specific kinds of problem explications and revisions/revision prescriptions result from developmental testing and debriefing?

Comprehension/recall kinds of problem explications were overwhelmingly the most frequent kind of response — or in other words problems relating to semantics. Student test subjects would often indicate, for example, that they did not understand something, there was too much information to remember, or that an example needed to be added. At times test subjects would only locate the problem without being able to explain their difficulty. Sometimes a subject would point to a perceived inaccuracy with the content or suggest a completely different medium from text. The physical features of the materials at times caused problems. Perceived problems in this category included a lack of highlighting of terms, and format and layout problems. Motivation and attention was another kind of problem explication generated by the sample. Irrelevant or redundant information or responses indicating that the material was not interesting were typical of responses in this category. The final kind of problem explication was a response which overlapped two or more of the above categories. A test subject might indicate that there was both a comprehension problem and a physical features problem in the same sentence, for example.

For revision, the categories were identical. Regardless of the kind of identified problem, test subjects generally generated revisions during instruction that would be classified as relating to the physical features of the materials, e.g., underline important terms, use point form, number the

sections, etc. It appears that test subjects typically generate revisions that are related more to the superficial aspects of the material than semantic aspects.

The specific kinds of problem explications and revisions can be found in Tables 72 and 73.

Do experimental conditions affect the amounts of general kinds of problem explications and revisions/revision prescriptions made during instruction?

Although the experimental conditions employed in the 2 x 2 ANOVA design detected differences in the amounts of data obtained during instruction, it is through the 2 x 2 and the 2 x 2 x 2 MANOVA designs which permit the investigation of differences in the kinds of data obtained during instruction.

The one-to-one active condition produced the greatest number of comprehension/recall kinds of problems. Test subjects in this condition appear to be less inhibited to reveal these important kinds of perceived problems during instruction than the other experimental conditions.

The small group active condition produced the fewest amounts of the kinds of feedback data. Since there was no difference in the proportions of respondents to the survey who used one-to-one and small group in combination as opposed to the exclusive use of small group as a general strategy, it should be noted that small group active might have a tendency to suppress test subjects comments.

The practitioners who responded to the survey indicated that test subjects do not actually revise the material. Test subjects are actively encouraged to indicate and comment on the problems they are encountering. Respondents to the survey strongly noted that after the developmental testing session is over, it is the exclusive responsibility of the developer/evaluator to revise based on the data provided through having developmentally tested the instruction. Developers probably revise intuitively based, in part, on the comments that test subjects make. The categorized and tabulated revision data from the experiment lends empirical support for a role that test subjects are able to play during developmental testing. Test subjects are better able at spotting perceived problems and explaining those problems to a developer/evaluator than they are at actually revising or suggesting revisions. When they do revise, test subjects attend to the physical features of the material — the more superficial aspects. No significant differences were detected across the experimental conditions for the kinds of revisions. It

can be concluded that test subjects, regardless of experimental condition, were equally poor at making revisions.

Do experimental conditions in conjunction with the test subject characteristic of prior knowledge of the material affect the amounts of general kinds of problem explications and revisions/revision prescriptions made during instruction?

Although more data can be obtained from test subjects with prior knowledge they do not produce more of one particular kind of problem explication or revision. Most of the respondents to the survey indicated that a mix of high, medium, and low level/aptitude test subjects should be used. Even if prior knowledge test subjects were to be included in a sample, developmental testing should attend to the data from members of the target population.

All of the results and conclusions presented based on the results of the reported research are meant to describe the process of developmental testing and debriefing rather than to judge which method is superior. Undoubtedly, the answers to the research questions addressed in this study raise more questions than had originally been asked. The final chapter will present some heuristics for developmental testing, based on both the results of the survey and the experiment. Some of the implications of the reported research for future research endeavors will be raised.

CHAPTER IX

**HEURISTICS FOR DESIGNING AND CONDUCTING DEVELOPMENTAL TESTING SESSIONS AND
RECOMMENDATIONS FOR FUTURE RESEARCH****Heuristics for Developmental Testing**

Based on the response patterns that emerged from the questionnaire which surveyed expert instructional designers' practices for designing and conducting developmental testing together with the results of the experiment, the following heuristics for developmental testing are offered as possibly useful recommendations for instructional designers.

Selecting the Test Subjects

From the target population, randomly select a "representative" sample of test subjects for whom the instruction has been designed.

If the target population is heterogeneous in terms of ability/aptitude, past achievement, motivational level, verbal level, or any other variable which might affect learning and performance then stratify the sample so that it reflects these characteristics, traits, etc. For example, if the target population can be stratified into high, medium, and low levels based on past achievement, then the sample of test subjects selected to developmentally test the materials should represent high, medium, and low level achievers. From each of these stratified groups, randomly select a test subject.

Select a recent graduate of the training/educational programme or course in which the material will be used. Feedback of a more critical nature might be generated. Pre-test results should indicate or confirm that this kind of test subject (a quasi-expert) has already attained the level of criterion performance specified.

One-to-One Testing and Debriefing

Conduct several one-to-one sessions, revising between sessions if necessary, and assume an active role as a developer/evaluator. The role of the test subject should be active, as well, to complement the developer/evaluator role.

Test out revisions made from one session during the next session.

Pretest the subject, administer the learning material, interact with the subject whenever a problem is perceived (by either the developer/evaluator or the subject) in order to locate and define the problem.

Systematically probe the subject to identify, explain, and revise possible problems and acknowledge that critical comments are being welcomed.

Give just enough help to get the subject back on track being aware that too much "tutoring" will contaminate the posttest results.

Do not insist that the test subject must revise or suggest a revision. Some students are better able than others to assume the role of reviser. Some subjects will only be able to pinpoint the location of where they perceive a problem to be. Ultimately, it is the developer/evaluator's role to make the revisions.

Posttest the subject, administer a formal attitudinal survey to the subject, and allow time for a short break. The subject should be posttested before the completion of the attitudinal survey as items on the survey should tap reactions to the posttest.

After having tallied the results from developmental testing, debrief the test subject based on: 1) their comments made during instruction (including any comments made while completing the self or posttest), 2) errors made on the self or posttest, and 3) unfavorable responses to the attitudinal survey. A test subject should also be informed that new problems, i.e., problems which were not detected during developmental testing, can also be raised.

Conduct an error analysis with the test subject, based on posttest results, in order to determine why the test subject made the particular error. Cooperatively analyze and interpret the tabulated data with the test subjects.

Revise, after the test subject has been debriefed, based on the test subject's behavior during the session. Then developmentally test with another test subject using the procedures outlined above.

Terminate one-to-one testing when the cognitive, psychomotor, and/or affective objectives are being satisfactorily met by high, medium, and low level test subjects (if the target population can be stratified according to known or presumed critical attributes).

Record all verbal interactions with the participants of testing sessions in both note and audio formats.

Small Group Testing

Developmentally test with small group(s) when one-to-one testing reveals that the objectives for the instruction are being met. Do not use a large number of test subjects during small group testing as it will be more difficult to manage larger numbers of subjects. The experimenter found that six test subjects was a manageable size for a group, especially for conducting the debriefings.

Assign a passive role to small group participants during instruction. Do not interact with the test subjects unless there is some critical problem that prevents a subject from completing the materials. Tell the test subjects where to find the tester if such an event occurs.

Ask the test subjects to write out their problem explications and suggestions on the materials as they might be intimidated to verbally reveal their learning difficulties.

Assign individual study areas to individual members of a group, if the instruction is truly designed to be self-instructional, in order to further reduce any possible intimidating peer influences as the subjects are trying to both learn from the material and criticize it at the same time. This would erase any possibility of intimidation. If the materials are self-instructional in nature but intended to be used within a classroom environment then having the test subjects sit together would simulate the way in which the materials will be eventually used.

Conduct the administration of the posttest so as to insure the integrity of the results.

After all subjects have completed the instructional materials, completed the test and filled out the attitudinal survey, ask them to reconvene for the debriefing.

Compile the debriefing agenda by: 1) listing all of the problems which the subjects identified during instruction (including problems identified while taking the achievement test), 2) identifying errors made on the posttest, and 3) identifying unfavorable responses to the attitudinal survey. It might be too restrictive to impose a criterion, for agenda setting purposes, on comments made during small group testing. Criteria might be applied to posttest errors and attitudinal survey results for small group agenda setting purposes; if no criteria are applied to these sources of feedback then perhaps all or a great number of the posttest and survey items will have to be discussed. This would expand the amount of time it would take to conduct a debriefing and risk boring or exhausting the test subjects. If criteria are utilized, then test subjects should be informed of them and the tester might reinforce that all feedback is valuable even if some of it is not being dealt with due to the imposed criteria. This should indicate to the group that individual feedback is not being "discounted"

During the debriefing of a small group, bring each item on the agenda up for discussion by the group as a whole. Do not tell the group who had indicated that there were problems while trying to learn from the materials and do not reveal to the group who made errors on the posttest or who had indicated dissatisfaction on the attitudinal survey. Guarantee the anonymity/ confidentiality of an individual test subject's response. Inform the test subjects that new problems, i.e., problems which were not detected during instruction, can also be raised for group discussion. Assume the role of session leader by presenting the agenda, keeping the subjects on task, encouraging comments, etc. Cooperatively analyze and interpret the data with the test subjects and attempt to get the group to reach a consensus for each item on the agenda. Tape record all verbal interactions with the participants of testing sessions. Transcribe this data after the session.

If necessary, revise after the test subject(s) have been debriefed based on the test subjects' behavior during the session. Test out any revisions during the next test session.

Terminate small group testing when a group meets the cognitive, psychomotor, and/or affective objectives.

Recommendations for Future Research

Preparing subjects

First, a technical matter: the preparation of subjects. In further research investigating developmental testing, it might be useful to avoid a major problem that was inherent to the experimental design which was utilized in this study. For this experimental study, subjects were tested over two days with two units of instructional material. Contaminating influences of this procedure have been discussed. One way to eliminate this source of contamination would be to have a day zero for purposes of training test subjects used in an experiment. This would insure that all subjects were cognizant of what was expected of them. It should be reiterated that placing student test subjects into the role of a critic is a not only demanding but also a relatively novel role for students. A day zero might utilize instruction that is totally unrelated to the subject matter to be used in a true or quasi-experimental design setting.

Other Media

The stimulus materials used in the reported experiment were self-instructional print materials. A considerable body of research exists for the film/television medium. Future proposed research could compare constraints on data gathering in various media including live instruction and compare the effects of a technique across media.

Other Populations

The population sampled for the Microbiology units which were used in the experiment were adult learners enrolled in a professional nursing programme at the junior college level. It would be useful to determine the kinds of data and the effectiveness of gathered data from other populations of learners such as kindergarten students, elementary school children, junior and high school students, the learning disabled, the emotionally disturbed, the retarded, the physically disabled, professionals taking continuing education courses, or high achieving groups of students. Within populations, it would be

extremely useful if abilities/aptitudes and personological variables could be identified which characterize or profile optimal test subjects.

Identifying Responders

Many of the test subjects used in the experiment were "non-responders" in that they did not generate comments (feedback data) during developmental testing. The purpose of developmental testing is to generate this kind of data. Perhaps by identifying, for selection purposes, "high-responders" expenditures of a great deal of time, resources, and effort utilized during formative evaluations might be reduced. Berthelot (1978) has conducted research on this problem using audio-visual materials in a tutorial setting and has produced a protocol for determining the verbal level of test subjects. Further research on the use of this protocol with different instructional media and other learner populations would be useful for establishing test subject selection procedures.

Revising Instruction

The focus of the reported research has been on the design and conduct of developmental testing and debriefing based on test results. Revision activities have not been the focus of this research. A logical extension of the research program would be to use the results from an experimental study such as the one just presented and revise based on the data gathered in order to determine which conditions produce more effective revisions. Purported algorithms and heuristics exist in the instructional design literature and they could be used in an experimental context. In this way, much needed empirical evidence could be gathered on the relative effectiveness of varying recommendations for revision activities. As well as generating empirical support for revision prescriptions, research conducted along these lines could be used to compare revisions across revisers for descriptive research purposes. The revised products could be content analyzed in order to ascertain the differential effects of role and number of participants during developmental testing on the process of revision. The relative effectiveness of revised instructional materials appears to be related to two critical revision variables. The first variable is how a reviser interprets formative evaluation data. The second variable is how a

reviser implements actual revisions based on the data. Research such as this would provide a foundation for better understanding the revision process.

On-line versus Retrospective Data Collection

In the area of cognitive science, there has been a long and frequent debate about the relative effectiveness of two very different ways to obtain data for experimental purposes. Ericsson and Simon (1984) have noted that it can be assumed that collection of retrospective data (recall protocols) after a delayed period of time from the presentation of stimuli will result in the distortion and loss of recalled information. This method of collecting data can be related to the debriefing strategy employed after instruction for the experiment which was just reported. It provided a considerable amount of unique feedback data. Usually, experiments investigating various aspects of comprehension have tended to employ "on-line" or process-related methods such as the probing of subjects used in experiments as they are attempting to learn, recall, or perform a given task. Ericsson and Simon (1984) suggest that it is through this think aloud method of collecting data that the loss and/or distortion of the data will be reduced. This data collection method can be related to the probing procedure used during one-to-one active experimental developmental testing condition. Having individual test subjects produce think aloud protocols did produce the greatest amount of verbalized comprehension/recall problems in the experiment. Ericsson and Simon's (1984) recommendations for collecting data from experimental settings are assumptions which have not been empirically tested. It should also be noted that there may be real differences between (a) having to read and think aloud and (b) having to learn new material and edit text concurrently.

Perhaps a better understanding of the effects of the role of the test subject and the developer / evaluator, as they interact with each other, can be gained from the different kinds of data collected during formative evaluation. One-to-one subjects could be assigned to a passive condition in which they are not even asked to identify, explicate, and remediate perceived problems during instruction. These subjects could be debriefed in order to obtain this kind of data. Other subjects could be assigned to a passive condition in which they are asked to identify, explicate, and remediate perceived problems

during instruction. The data from the two groups could be compared. Data from these two groups could also be compared to an active condition in which both the experimenter and the subject assume active roles for the purposes of obtaining on-line protocol data using probing techniques. Results of this kind of research would be useful to a number of research communities.

Embedded Problems

The stimulus materials which were used in the experiment had been produced by teams of experts following the guidelines for systematic instructional design. These materials were not contrived for experimental purposes. Any problems inherent within these materials resulted during their development in the field. It might be useful to replicate the study which has been reported using contrived materials which have embedded "problems" placed in them for experimental purposes. By intentionally violating principles of instructional design, resulting in hypothetical design errors, stronger empirical conclusions could be made about the ability of target learners to identify, explicate, and revise instruction. This would lend empirical support to the numerous design principles being espoused. This kind of research might also result, eventually, in transforming instructional design models (which are presented as design heuristics) to the point of genesis of a valid, general theory of instruction.

There is a need to conduct programmatic empirical research into the validation process of instruction. Some of the variables in the process of formative evaluation have been examined here. Future research will undoubtedly benefit those who design instruction systematically.

References

- Abedor, A. (1971). Development and validation of a model explicating the formative evaluation process for multi-media self-instructional learning systems (Doctoral dissertation, Michigan State University, Lansing, MI, 1971). Dissertation Abstracts International, 32, 4915A.
- Andrews, D., & Goodson, L. (1980). A comparative analysis of models of instructional design. Journal of Instructional Development, 3(4), 2 - 16.
- Baggaley, J. (1986). Formative evaluation of educational television. Canadian Journal of Educational Communication, 15(1), 29 - 43.
- Baghdadi, A. (1980). A comparison between two formative evaluation methods (Doctoral dissertation, Indiana University, 1980). Dissertation Abstracts International, 41, 3387A.
- Baker, E. (1970, March). An experimental appraisal of the generalizability of empirical revision procedures for curriculum materials. Paper presented at the Annual Meeting of the American Research Association, Minneapolis, MN.
- Baker, E. (1972). Preparing instructional materials for educational developers. Final Report project No. 1-0027, U.S. Dept. of H.E.W., Office of Education, National Center for Educational Research and Development.
- Baker, E. (1973). The technology of instructional development. In R. Travers (Ed.), Second handbook of research on teaching. American Educational Research Association, Chicago, IL: Rand McNally.
- Baker, E. (1974). Formative evaluation of instruction. In W. James Popham (Ed.), Evaluation in education: Current Applications. American Educational Research Association, Berkeley, CA: McCutchan Publishing Company.
- Baker, E., & Alkin, M. (1973). Formative evaluation in instructional development. AV Communication Review, 21(4), 389 - 418.
- Baker, E., & Schutz, R. (1971). Instructional product development. New York: Van Nostrand Reinhold.
- Bell, N., & Abedor, A. (1977). Developing audio-visual instructional modules for vocational and technical training. Englewood Cliffs, N.J.: Educational Technology Publications.
- Berthelot, S. (1978). Le niveau de verbalisation d'un sujet dans le cadre de l'application d'une evaluation LYR d'un document audiovisuel par le mode tutoriel. Unpublished doctoral dissertation. Universite de Montreal.
- Bjersstedt, A. (1972). Educational technology: Instructional programming and didakometry. New York: McGraw Hill.
- Bloom, B., Engelhart, E., Faust, E., Hill, W., and Krathwohl, D. (1956) Taxonomy of educational objectives. Handbook I: Cognitive domain. New York: David McKay, Inc.

- Brandenburg, D. (1982). Some improvements for the ubiquitous evaluation questionnaire. NSPI Journal, October, 9 - 11.
- Bratton, B. (1981). Training the instructional development specialist to work in unfamiliar content areas. Journal of Instructional development, 4(3), 21-23.
- Brethower, K. (1965). Individual testing as a guide to revision. In: Selected N.S.P.I. Talks, 1965: Occasional Paper #5. The University of Michigan, Ann Arbor: Center for Programmed Learning for Business, 39 - 44.
- Briggs, L. (1970). Handbook of procedures for the design of instruction. Pittsburgh, PA: American Institutes for Research.
- Burkholder, B. (1982). The effectiveness of using the instructional strategy diagnostic profile to prescribe improvements in self-instructional materials teaching abstract concepts. Journal of Instructional Development, 5(2), 2 - 9.
- Burt, C., & Geis, G. (1986, April). Guidelines for developmental testing: Proposed and practiced. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. (ERIC Documentation Reproduction Service No. ED 270 500)
- Burt, C. (1987, April). An experimental study of group size and participants' role in developmental testing. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC. (ERIC Documentation Reproduction Service No. ED 292 853)
- Cambre, M. (1981). Historical overview of formative evaluation of instructional media products. Educational Communication and Technology Journal, 29(1), 3 - 25.
- Choi, S., Merrill, D., Callahan, E., Hawkins, R., & Norton, R. (1979). The relationship of test performance to ISDP rating in organic chemistry texts. Journal of Instructional Development, 3(1), 16 - 25.
- Clarke, J. (1975). A model for pre-trial formative curriculum evaluation: Its development and application. Australian Journal of Education, 19(3), 251 - 267.
- Corey, G. & Corey, M. (1982). Groups: Processes and practice (2nd ed.) Belmont, CA: Wadsworth.
- Cunningham, D. (1971). Formative evaluation of replicable forms. Paper presented at the annual meeting of the American Educational Research Association, New York.
- Davis, R., Alexander, L., & Yelon, S. (1974). Learning system design: An approach to the improvement of instruction. New York: McGraw-Hill.
- Debert, R. (1979). Revisionary tactics. National Society for Performance and Instruction Journal, 18(7), 18 - 21.
- Deterline, W., & Lenn, P. (1972). Coordinated instructional systems. Palo Alto, CA: Sound Education, Inc.
- Dick, W. (1968). A methodology for the formative evaluation of instructional materials. Journal of Educational Measurement, 5(2), 99 - 102.

- Dick, W. (1977). Formative evaluation. In Briggs, L. J. (Ed.), Instructional design: Principles and applications. Englewood Cliffs N.J.: Educational Technology Publications.
- Dick, W. (1979, March). Current research and issues in formative evaluation. Paper presented at the Annual Meeting of the Association of Educational Communications and technology, New Orleans, LA.
- Dick, W. (1980). Formative evaluation in instructional development. Journal of Instructional Development, 3(3), 3 - 60.
- Dick, W. (1986, April). An analysis of a sample of the learner verification and revision reports submitted to the Florida Department of Education in 1985 - 86. Paper presented at the annual meeting of the American Education Research Association, San Francisco, CA.
- Dick, W. (1987, April). A large-scale application of formative evaluation of instructional material. Paper presented at the annual meeting of the American Education Research Association, Washington, D.C..
- Dick, W. & Carey, L. (1978). The systematic design of instruction. Glenview, IL.: Scott, Foresman & Company.
- Dick, W. & Carey, L. (1985). The systematic design of instruction (2nd ed.). Glenview, IL.: Scott, Foresman & Company.
- Dupont, D. (1977). Etude comparative de deux facons de recueillir les informations dans le cadre d'une evaluation LVR d'un document audiovisuel. Universite de Montreal. Unpublished Master's thesis.
- Dupont, D. (1980). Developpement et evaluation d'un modele de revision de documents audio-scripto-visuels destine aux reviseurs. GRERDAVE, Rapport 17. Montreal, Quebec: Universite de Montreal.
- Dupont, D., & Stolovitch, H. (1983, March). The effects of a systematic revision model on revisers in terms of student outcomes. National Society for Performance and Instruction Journal, 33 - 37.
- Englemann, Z. (1983, March). Theory of instruction. National Society for Performance and Instruction Journal, 13 - 16.
- Engler, D. (1976). Learner verification: A fine-grain analysis of go-go empiricism. Audiovisual Communication Review, 24(1), 5 - 20.
- EPIE Institute (1975). Pilot guidelines for improving instructional materials through the process of learner verification and revision. New York: Author.
- Ericsson, K. & Simon, H. (1984). Protocol analysis: Verbal reports as data. Cambridge, MA: The MIT Press.
- Feer, R. (1973). The evaluation interview (2nd ed.). New York: McGraw Hill.
- Finn, J. & Mattsson, I. (1978). Multivariate analysis in educational research: Applications of the MULTIVARIANCE program. Chicago, Ill: National Education Resource.

- Frase, L., DeGracie, J., & Poston, W. (1974). Product validation: Pilot test or panel review? Educational Technology, 14(8), 32 - 35.
- Friesen, P. (1973). Designing instruction. Santa Monica, CA: Miller Publishing Company.
- Gagne, R., & Briggs, L. (1979). Principles of instructional design (2nd ed.). New York: Holt, Rinehart, & Winston.
- Gagnon, C., & Stolovitch, H. (1980). Etude de deux documents pedagogiques de francais au niveau elementaire 1er cycle, Rapport GRERDAVE, No. 17, Universite de Montreal.
- Gallagher, J. (1984). The modification and evaluation of an instructional design model which includes adult learner involvement (Doctoral dissertation, Florida State University, 1984). Dissertation Abstracts International, 45, 1246A.
- Geis, G. (1973). A comment on validated instruction. Improving Human Performance, 3, 162-163.
- Geis, G. (1986, April). Student input in materials development. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.
- Geis, G., Burt, C., & Weston, C. (1984, April). Instructional development: Developmental testing. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 243 793)
- Geis, G., & Weston, C. (1983). Performance system design. Unpublished manuscript, McGill University, Montreal.
- Glass, G. & Hopkins, K. (1984). Statistical methods in education and psychology (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Golas, K. (1983). Formative evaluation effectiveness and cost: Alternative models for evaluating printed instructional materials. Performance and Instruction Journal, 22(5), 17 - 19.
- Gropper, G., Lumsdaine, A., & Shipman, V. (1961). Improvement of televised instruction based on student responses to achievement tests. Pittsburgh, PA: American Institutes for Research, Report No. 1, Studies in Televised Instruction.
- Gropper, G. (1975). Diagnosis and revision in the development of instructional materials. Englewood Cliffs, NJ: Educational Technology Publications.
- Hartley, J., & Burnhill, P. (1977). Fifty guidelines for improving instructional text. Programmed Learning and Educational Technology, 14(1), 65 - 74.
- Henderson, E., & Nathenson, M. (1977). Case study in the implementation of innovation: A new model for developmental testing. In Philip Hills and John Gilbert (Eds.) Aspects of educational technology: Volume X The spread of educational technology. London: Kogan Page.
- Holsti, O. (1968). Content analysis for the social sciences and humanities. Reading, MA: Addison-Wesley.
- Holsti, O. (1969). Content analysis. In G. Lindzey and E. Aronson (Eds.), The handbook of social psychology. Reading, MA: Addison-Wesley.

- Horn, R. (1964). What programming errors can be discovered by student testing? Programmed Instruction, 4(2), 6 - 11.
- Horn, R. (1964). Developmental testing: Trying out programmed instructional materials with individual students. University of Michigan, Ann Arbor, MI: The Center for Programmed Learning for Business.
- Horn, R. (1966). Developmental testing. Ann Arbor, MI: Center for Programmed Learning for Business.
- Horn, R. (1976). How to write information mapping. Lexington, MA: Information Resources Inc., P.O. Box 417.
- Jacobs, P. (1962a). Some implications of testing procedures for auto-instructional programming. Contract No. AF33(616) - 7795: Aerospace Medical Research Laboratories, Project No. 1710, Task No. 171007.
- Jacobs, P. (1962b). Some relationships between testing and auto-instruction programming. Audiovisual Communication Review, 10(6), 317 - 327.
- Johnson, R., & Johnson, S. (1975). Toward individualized learning: A developer's guide to self-instruction. Reading, Massachusetts: Addison-Wesley Publishing Company.
- Kandaswamy, S. (1980). Evaluation of instructional materials: A synthesis of models and methods. Educational Technology, 20(6), 19 - 26.
- Kandaswamy, S., Stolovitch, H., & Thiagarajan, S. (1976). Learner verification and revision: An experimental comparison of two methods. Audiovisual Communication Review, 24(3), 316 - 328.
- Keppel, G. (1982). Design and analysis: A researcher's handbook (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Knirk, F., & Gustafson, K. (1986). Instructional technology: A systematic approach to education. New York: Holt, Rinehart and Winston.
- Komoski, K. (1974). An imbalance of product quantity and instructional quality: The imperative of empiricism. Audiovisual Communication Review, 22(4), 357 - 386.
- Komoski, P. (1976). Pilot guidelines for improving instructional materials through the process of learner verification and revision. New York: E.P.I.E. Institute.
- Komoski, P. K., & Woodward, A. (1985). The continuing need for the learner verification and revision of textual material. In David H. Jonassen (Ed.), The technology of text, Volume Two: Principles for structuring, designing, and displaying text. Englewood Cliffs, N. J.: Educational Technology Publications.
- Lawson, T. (1974). Formative instructional product evaluation: Instruments and strategies. Englewood Cliffs, NJ: Educational Technology Publications.
- Levin, H. (1974). A conceptual framework for accountability in education. School Review, 82, 363 - 391.

- Lumsdaine, A. (1965). Assessing the effectiveness of instructional programs. In R. Glaser (Ed.), Teaching machines and programmed learning, II: Data and directions. Washington, DC: National Education Association.
- Merkle, D. (1967). Final report: The development of the Bell System first aid and personal safety course - An exercise in the application of empirical methods to instructional systems design. Palo Alto, CA: American Institutes for Research in Behavioral Sciences. (ERIC Document Reproduction Service No. ED 026 871)
- Merkle, S. M. (1967). Empirical testing of programs. In P. C. Lange (Ed.), Programmed Instruction. The Sixty-sixth Yearbook of the National Society for the Study of Education, Part II. Chicago, IL: NSSE.
- Merkle, S. (1978). Designs for instructional designers. Champaign, IL: Stipes Publishing Company.
- Merkle, S. (1983). Book review. Performance and Instruction Journal, 22(5), 28 - 29.
- McCrary, R. D. (1964). The design method - A scientific approach to valid design. In J. R. Bright (Ed.), Research, development, and technological innovation - An introduction. Homewood, IL: Richard D. Irwin, 59 - 67.
- McGrath, J., & Altman, I. (1966). Small group research: A synthesis and critique of the field. New York: Holt, Rinehart, & Winston.
- Merrill, D. M. & Boutwell, R. C. (1973). Instructional development: Methodology and research. In F. N. Kerlinger (Ed.) Review of Research in Education. Itasca, IL: F.E. Peacock.
- Merrill, D., & Tennyson, R. (1977). Teaching concepts: An instructional design guide. Englewood Cliffs, NJ: Educational Technology Publications.
- Merrill, D., Reigeluth, C., & Faust, G. (1979). The instructional quality profile: A curriculum evaluation and design tool. In H. O'Neill (Ed.), Procedures for instructional systems development. New York: Academic Press.
- Merrill, M. D., Richards, R. E., Schmidt, R. V., & Wood, N. D. (1977). The instructional strategy diagnostic profile: Training manual. Orem, UT: Courseware, Inc.
- Montague, W., Ellis, J., & Wulfeck, W. (1983). Instructional quality inventory: A formative evaluation tool for instructional development. Performance and Instruction Journal, 22(5), 11 - 13.
- Nathenson, M., & Henderson, E. (1977). Problems and issues in developmental testing. National Society for Performance and Instruction Journal, 16(1), 9 - 10.
- Nathenson, M., & Henderson, E. (1979). Using student feedback to improve learning materials. London: Croom Helm.
- Newman, S. (1957). Student vs. instructor design of study method. Journal of Educational Psychology, 48(6), 328 - 333.
- Passmore, D., Asche, J., & Chase, E. (1975). Delineation of information for formative evaluation of educational products. Paper presented at the meeting of the New England Educational Research Association.

- Poulson, C. (1969). Evaluation of instructional systems. In J. Crawford (Ed.), CORD: National research training manual. Teaching Research Division: Oregon State System of Higher Education.
- Popper, K. R. (1959). The logic of scientific discovery. London: Hutchinson.
- Popper, K. R. (1973). Objective knowledge: An evolutionary approach. Oxford: Oxford University Press.
- Reigeluth, C. (Ed.). (1983). Instructional design theories and models: An overview of their current status. Hillsdale, NJ: Lawrence Erlbaum.
- Reigeluth, C., Van Patten, J., & Doughty, P. (1981). Science approach to instructional development. Performance and Instruction Journal, 20(7), 19 - 22.
- Repicky, P., & Hotvedt, M. (1977). The role of formative evaluation in a continuing education program: Basic rehabilitation nursing techniques. The Journal of Continuing Education in Nursing, 8(4), 24 - 34.
- Robeck, M. (1965). A study of the revision process in programmed instruction. Unpublished master's thesis, University of California, Los Angeles.
- Robinson, T. (1972). Replicable training in revision techniques. Dissertation Abstracts International, 33, 04A. (University Microfilms No. 72-25, 826)
- Rodriguez, A., & Rodriguez, C. (1982). Application de l'approche L.V.R. a differents stades de la realisation de materiel d'enseignement audio-visuel. Unpublished master's thesis, University of Montreal.
- Rose, N., & Van Horn, C. (1956). Theory and application of preproduction testing. AV Communication Review, 4, 21 - 30.
- Rosen, M. (1968). Final report: An experimental design for comparing the effects of instructional media programming procedures. Palo Alto, CA: American Institutes for Research in Behavioral Sciences. Sponsored by the U. S. Department of Health, Education, and Welfare, Project No 5 - 8332.
- Rothkopf, E. (1963). Some observations on predicting instructional effectiveness by simple inspection. The Journal of Programmed Instruction, 2, 19 - 20.
- Sanders, J., & Cunningham, D. (1973). A structure for formative evaluation in product development. Review of Educational Research, 43(2), 217 - 236.
- Sanders, J., & Cunningham, D. (1974). Techniques and procedures for formative evaluation. Portland, OR: Northwest Regional Educational Laboratory.
- Sarapin, M. (1981). Formative evaluation: A model for instructional material development and revision. Journal of Industrial Teacher Education, 18(3), 3 - 13.
- Scott, R., & Yelon, S. (1969, October). The student as a co-author: The first step in formative evaluation. Educational Technology, 76 - 78.

- Scriven, M. (1967). The methodology of evaluation. In Tyler, R.W. et al. Perspectives of curriculum evaluation. (AERA Monograph Series on Curriculum Evaluation, No. 1), Chicago, IL: Rand McNally.
- Sedlik, J. (1971). Systems engineering of education XIV: Systems techniques for pretesting mediated instructional materials. Los Angeles, CA: Education and Training Consultants Co.
- Smith, K. (1978). The utilization of teacher feedback in ETV development. Educational Technology, 18(19), 49 - 51.
- Smith, N., & Murray, S. (1975, March). The status of research on models of product development and evaluation. Educational Technology, 13 - 17.
- State of Florida Department of Education (1986). Guidelines for the learner-verification and revision process. Tallahassee, FL: State Instructional Materials Councils.
- State of Florida (1986). Module V. Learner verification and review process. Tallahassee, FL: Department of State.
- Stolovitch, H. (1982). Applications of the intermediate technology of learner verification and revision (LVR) for adapting international instructional resources to meet local needs. Performance and Instruction Journal, 21(7), 16 - 22.
- Stolovitch, H., & Berthelot, S. (1982, April). Rapid verification and revision of the readability and comprehensibility levels of instructional text. Paper presented at the twenty-first annual conference of the National Society for Performance and Instruction, San Diego, CA.
- Thiagarajan, S. (1976). Learner verification and revision: What, who, when, and how? Audiovisual Instruction, 21(1), 18 - 19.
- Thiagarajan, S. (1978). Instructional product verification and revision: 20 questions and 200 speculations. Educational Communications and Technology Journal, 26(2), 133 - 141.
- Thiagarajan, S. (1986). Debriefing. Performance and Instruction, 25(5), 45.
- Thiagarajan, S., Semmel, D., & Semmel, M. (1974). Instructional development for training teachers of exceptional children: A sourcebook. Bloomington, IN: Center for Innovation in Teaching Handicapped.
- Tyler, R. (1942). General statement on evaluation. Journal of Educational Research, 35, 492 - 501.
- VanderMeer, A., Thorne, H., & Montgomery, R. (1961). An investigation of the improvement of informational filmstrips and the derivation of principles relating to the effectiveness of these media: Studies I - III. Abstracted in VanderMeer, A. Ten recent Title VII Reports: A perspective. Audiovisual Communication Review, 9(6).
- Vedros, R. (1986, April). Improving textbooks: LVR Florida style. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Wager, J. (1983). One-to-one and small group formative evaluation. Performance and Instruction Journal, 22(5), 5 - 7.

- Warwick, D. & Leninger, C. (1975). The sample survey: Theory and practice. New York: McGraw Hill.
- Weston, C. (1986). Formative evaluation of instructional materials: An overview of approaches. Canadian Journal of Educational Communication, 15(1), 5 - 17.
- Weston, C. (1987a). The importance of involving experts and learners in formative evaluation. Canadian Journal of Educational Communication, 16(1), 45 - 58.
- Weston, C. (1987b, April). A controlled application of developmental testing and expert review. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.
- Weston, C., Burt, C., & Geis, G. (1984, April). Instructional development: Revision procedures. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Wiener, N. (1954). The human use of human beings. Garden City, NY: Doubleday and Co., Inc.
- Wile, M. (1979). Educational product evaluation in the health sciences: Formative and summative components. Educational Technology, 19(1), 52 - 57.
- Wood, N. D., Richards, R. E., and Merrill, M. D. (1976). Prediction of student performance on rule using tasks from the diagnosis of instructional strategies. Provo, UT: Brigham Young University Research Paper.
- Yelon, S. (1974). Constructive evaluation: Improving large scale instructional projects. Lansing, MI.

Appendix A

References Reviewed for Purposes of Identifying Guidelines

- Abedor, A. (1971). Development and validation of a model explicating the formative evaluation process for multi-media self-instructional learning systems (Doctoral dissertation, Michigan State University, Lansing, MI, 1971). Dissertation Abstracts International, 32, 4915A.
- Baker, E., & Schutz, R. (1971). Instructional product development. New York: Van Nostrand Reinhold.
- Bell, N., & Abedor, A. (1977). Developing audio-visual instructional modules for vocational and technical training. Englewood Cliffs, NJ: Educational Technology Publications.
- Björstedt, A. (1972). Educational technology: Instructional programming and didakometry. New York: McGraw Hill.
- Bloom, B., Hastings, T., & Maudus, G. (1971). Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill.
- Briggs, L. (1967). Instructional media: A procedure for the design of multi-media instruction, a critical review of research, and suggestions for future research. Pittsburgh, PA: American Institutes for Research.
- Briggs, L., & Wager, W. (1981). Handbook of procedures for the design of instruction. Englewood Cliffs, NJ: Educational Technology Publications.
- Davis, R., Alexander, L., & Yelon, S. (1974). Learning system design. An approach to the improvement of instruction. New York: McGraw-Hill.
- Deterline, W., & Lenn, P. (1972). Coordinated instructional systems. Palo Alto, CA: Sound Education, Inc.
- Dick, W., & Carey, L. (1978). The systematic design of instruction. Glenview, IL: Scott, Foresman & Company.
- Friesen, P. (1973). Designing instruction. Santa Monica, CA: Miller Publishing Company.
- Gagne, R., & Briggs, L. (1979). Principles of instructional design (2nd ed.). New York: Holt, Rinehart, & Winston.
- Gropper, G. (1975). Diagnosis and revision in the development of instructional materials. Englewood Cliffs, NJ: Educational Technology Publications.
- Hartley, J. (1978). Designing instructional text. New York: Nichols Publishing Company.
- Holsclaw, J. (1974). The development of procedural guidelines for the systematic design of instruction within higher education (Doctoral dissertation, University of Southern California, 1974). Dissertation Abstracts International, 35, 1873A.

- Holtby, B. (1967). What the training officer needs to know about contracting for programs. In G. Rummier, J. Yaney, & A. Schrader (Eds.), Managing the instructional programming effort. University of Michigan: Bureau of Industrial Relations.
- Horn, R. (1966). Developmental testing. Ann Arbor, MI: Center for Programmed Learning for Business.
- Johnson, R., & Johnson, S. (1971). Assuring learning with self-instructional packages, or up the up staircase. Chapel Hill, NC: Self Instructional Packages, Inc.
- Johnson, R., & Johnson, S. (1975). Toward individualized learning: A developer's guide to self-instruction. Reading, MA: Addison-Wesley Publishing Company
- Laird, D. (1978). Approaches to training and development. Reading, MA: Addison-Wesley Publishing Company.
- Lawson, T. (1974). Formative instructional product evaluation: Instruments and strategies. Englewood Cliffs, NJ: Educational Technology Publications.
- Merkle, S. M. (1967). Empirical testing of programs. In P. C. Lange (Ed.), Programmed Instruction. The Sixty-sixth Yearbook of the National Society for the Study of Education, Part II. Chicago, IL: NSSE.
- Merkle, S. (1978) Designs for instructional designers. Champaign, IL. Stipes Publishing Company.
- Merrill, D., Reigeluth, C., & Faust, G. (1979). The instructional quality profile: A curriculum evaluation and design tool. In H. O'Neill (Ed.), Procedures for instructional systems development. New York: Academic Press.
- Merrill, D., & Tennyson, R. (1977). Teaching concepts: An instructional design guide. Englewood Cliffs, NJ: Educational Technology Publications.
- Nathenson, M., & Henderson, C. (1979). Using student feedback to improve learning materials. London: Croom Helm.
- Popham, J., & Baker, E. (1970) Systematic instruction. New Jersey: Prentice-Hall, Inc.
- Robeck, M. (1965). A study of the revision process in programmed instruction. Unpublished master's thesis, University of California, Los Angeles, CA.
- Robinson, T. (1972). Replicable training in revision techniques (Doctoral dissertation, University of California, Los Angeles, CA, 1972) Dissertation Abstracts International, 33, 04A.
- Rosen, M. (1968). Final report: An experimental design for comparing the effects of instructional media programming procedures. Palo Alto, CA: American Institutes for Research in Behavioral Sciences. Sponsored by the U. S. Department of Health, Education, and Welfare, Project No. 5 - 8332.
- Siebert, I. (1975). Educational technology: A handbook of standard terminology and guide for recording and reporting information about educational technology. Washington: U.S. Government Printing Office.

Thiagarajan, S., Semmel, D., & Semmel, M. (1974). Instructional development for training teachers of exceptional children: A sourcebook. Bloomington, IN: Center for Innovation in Teaching Handicapped.

West, C. & Foster, S. (1976). The psychology of human learning and instruction in education. Belmont, CA: Wadsworth Publishing Company.

Wong, M., & Raulerson, J. (1974). A guide to systematic instructional design. Englewood Cliffs, NJ: Educational Technology Publications.

Wydra, F. (1980). Learner controlled instruction. Englewood Cliffs, NJ: Educational Technology Publications.

Yelon, S. (1974). Constructive evaluation: Improving large scale instructional projects. Lansing, MI.

Appendix B

Selected Instructional Design Texts

- Abedor, A. (1971). Development and validation of a model explicating the formative evaluation process for multi-media self-instructional learning systems (Doctoral dissertation, Michigan State University, Lansing, MI, 1971). Dissertation Abstracts International, 32, 4915A.
- Baker, E., & Schutz, R. (1971). Instructional product development. New York: Van Nostrand Reinhold.
- Bell, N., & Abedor, A. (1977). Developing audio-visual instructional modules for vocational and technical training. Englewood Cliffs, NJ: Educational Technology Publications.
- Bjerstedt, A. (1972). Educational technology: Instructional programming and didakometry. New York: McGraw Hill.
- Davis, R., Alexander, L., & Yelon, S. (1974). Learning system design: An approach to the improvement of instruction. New York: McGraw-Hill.
- Deterline, W., & Lenn, P. (1972). Coordinated instructional systems. Palo Alto, CA: Sound Education, Inc.
- Dick, W., & Carey, L. (1978). The systematic design of instruction. Glenview, IL.: Scott, Foresman & Company.
- Friesen, P. (1973). Designing instruction. Santa Monica, CA: Miller Publishing Company.
- Gropper, G. (1975). Diagnosis and revision in the development of instructional materials. Englewood Cliffs, NJ: Educational Technology Publications.
- Horn, R. (1966). Developmental testing. Ann Arbor, MI: Center for Programmed Learning for Business.
- Johnson, R., & Johnson, S. (1975). Toward individualized learning: A developer's guide to self-instruction. Reading, MA: Addison-Wesley Publishing Company.
- Lawson, T. (1974). Formative instructional product evaluation: Instruments and strategies. Englewood Cliffs, NJ: Educational Technology Publications.
- Markle, S. (1978). Designs for instructional designers. Champaign, IL: Stipes Publishing Company.
- Merrill, D., & Tennyson, R. (1977). Teaching concepts: An instructional design guide. Englewood Cliffs, NJ: Educational Technology Publications.
- Thiagarajan, S., Semmel, D., & Semmel, M. (1974). Instructional development for training teachers of exceptional children: A sourcebook. Bloomington, IN: Center for Innovation in Teaching Handicapped.

APPENDIX C

Questionnaire on Developmental
Testing and Revision

Please return this questionnaire in the
self-addressed envelope enclosed.

c/o Dr. George Geis
Dept. of Educational Psyc.,
Faculty of Education
3700 McTavish St.
McGill University
Montreal, Quebec, Canada
H3A 1Y2

1. In developmentally testing proto-type instructional materials, which of the following general strategies do you usually employ? PLEASE CHECK, ☒ , THE MOST APPROPRIATE RESPONSE(S).

Frequency:

- 52 — ☐ testing with subjects on a one-to-one basis.
- 60 — ☐ testing with subjects on a small group basis.
- 58 — ☐ in addition to testing subjects, I obtain information from subject matter experts or peer reviews.
- 8 — ☐ instead of testing subjects, I obtain information from subject matter or peer reviews.
- 6 — ☐ OTHER (please outline briefly)

If any answers apply solely or primarily to one stage of developmental testing, indicate so by using this code (which will appear often in the questionnaire). Place the appropriate letter on the line next to your check mark(s).

- a. early stage
b. middle stage
c. late stage (just before field testing)

BASED ON EXPERIENCE, RESEARCH, OR ADVICE, WHAT HAS PROMPTED YOU TO UTILIZE YOUR SELECTION(S)?

2. In revising prototype instructional materials, which of the following general strategies do you usually employ? PLEASE CHECK, ☒, THE MOST APPROPRIATE RESPONSE(S).

Frequency:

- 61 — ☐ revision is based on test data.
- 22 — ☐ revision is based on an attitudinal survey.
- 69 — ☐ revision is based on test subjects' comments.
- 28 — ☐ revision takes place during the testing session.
- 70 — ☐ revision takes place after the testing session.
- 8 — ☐ OTHER (please outline briefly)

If any answers apply solely or primarily to one stage of developmental testing, indicate so by using this code (which will appear often in the questionnaire). Place the appropriate letter on the line next to your check mark(s).

- a. early stage
- b. middle stage
- c. late stage (just before field testing)

**BASED ON EXPERIENCE,
RESEARCH, OR ADVICE, WHAT
HAS PROMPTED YOU TO
UTILIZE YOUR SELECTION(S)?**

3. What aptitudes or levels of performance do you require of test subjects utilized during the developmental testing and revision phase of the instructional design process? PLEASE CHECK, ☒, THE MOST APPROPRIATE RESPONSE(S).

Frequency:

3

☐ high level/aptitude performers only.

11

☐ high and low level/aptitude performers.

52

☐ high, medium, and low level/aptitude performers.

11

☐ enthusiastic and highly verbal test subjects.

30

☐ OTHER Please outline briefly)

BASED ON EXPERIENCE,
RESEARCH, OR ADVICE, WHAT
HAS PROMPTED YOU TO
UTILIZE YOUR SELECTION(S)?



4. What kind of instrument do you usually use to differentiate test subject level/aptitude before the testing session? PLEASE CHECK, ☒, THE MOST APPROPRIATE RESPONSE(S).

Frequency:

- 3 ☐ I.Q. Test
- 5 ☐ a standardized aptitude test.
- 19 ☐ past achievement data (e.g., G.P.A.,
course grades, etc.).
- 32 ☐ job performance measures.
- 11 ☐ previous instructors' opinions.
- 30 ☐ pre-test on material to be taught.
- 39 ☐ OTHER (please outline briefly)
-
-
-
-
-

BASED ON EXPERIENCE,
RESEARCH, OR ADVICE, WHAT
HAS PROMPTED YOU TO
UTILIZE YOUR SELECTION(S)?



5. How many sessions and test subjects do you typically involve during developmental testing and revision? PLEASE FILL IN THE BLANK(S).

a) number of one-on-one sessions. _____

b) number of small group sessions. _____

c) number of test subjects typically used in each small group session. _____

OTHER (please outline briefly)

**BASED ON EXPERIENCE,
RESEARCH, OR ADVICE, WHAT
HAS PROMPTED YOU TO
UTILIZE YOUR SELECTION(S) ?**

What do you consider the role of the test subject to be during developmental testing?
PLEASE CHECK, ☒, THE MOST APPROPRIATE RESPONSE(S).

Frequency:

39 — ☐ test subject provides pre- and post-test
criterion data.

41 — ☐ test subject provides attitudinal data.

81 — ☐ test subject comments on problems with the
instructional material.

52 — ☐ test subject comments on problems with the
tests.

11 — ☐ OTHER (please outline briefly)

If any answers apply solely or primarily to one stage of developmental testing, indicate so by using this code (which will appear often in the questionnaire). Place the appropriate letter on the line next to your check mark(s).

- a. early stage
- b. middle stage
- c. late stage (just before field testing)

**BASED ON EXPERIENCE,
RESEARCH, OR ADVICE, WHAT
HAS PROMPTED YOU TO
UTILIZE YOUR SELECTION(S)?**

- a. early stage
- b. middle stage
- c. late stage (just before field testing)

- Frequency:**

- [illegible]

**BASED ON EXPERIENCE,
RESEARCH, OR ADVICE, WHAT
HAS PROMPTED YOU TO
UTILIZE YOUR SELECTION(S) ?**

8. Do you see different test subjects as contributing qualitatively different information during successive developmental testing and revision sessions or do you find that test subjects' comments generally replicate each other.

Frequency:

62

☐ different subjects often provide quite different kinds of information.

32

☐ subjects' generally replicate each others comments.

4

☐ OTHER (please outline briefly)

BASED ON EXPERIENCE,
RESEARCH, OR ADVICE, WHAT
HAS PROMPTED YOU TO
UTILIZE YOUR SELECTION(S)?



9. During a developmental testing session, test subjects are: PLEASE CHECK, ☒, THE MOST APPROPRIATE RESPONSE(S).

Frequency:

- 4 — ☐ not permitted to ask any questions.
- 2 — ☐ not permitted to make comments.
- 13 — ☐ permitted to but not encouraged to make comments or ask questions.
- 63 — ☐ actually encouraged to ask questions or make comments.
- 26 — ☐ actually taught by the developer when they run into difficulty.
- 37 — ☐ thinking aloud and requesting assistance when required.
- 47 — ☐ recording problems on a revision sheet and/or on the actual instructional materials as they work through the material.
- 6 — ☐ OTHER (please outline briefly)

If my answers apply solely or primarily to one stage of developmental testing, indicate so by using this code (which will appear often in the questionnaire). Place the appropriate letter on the line next to your check mark(s).

- a. early stage
b. middle stage
c. late stage (just before field testing)

BASED ON EXPERIENCE, RESEARCH, OR ADVICE, WHAT HAS PROMPTED YOU TO UTILIZE YOUR SELECTION(S)?

to be during developmental testing? PLEASE
CHECK, ☒, THE MOST APPROPRIATE RESPONSE(S).

Frequency:

- 56 — ☐ the developer waits, watches, makes notes, and keeps quiet while the test subject works.
- 68 — ☐ if a problem arises, the developer asks "why", pinpoints the problem, and gives just enough help to get the test subject back on track.
- 19 — ☐ the developer personally answers test subject's questions in a tutorial fashion.
- 29 — ☐ the developer notices gaps, predicts confusions, senses response failures, and re-writes items on the spot.
- 4 — ☐ OTHER (please outline briefly)

If any answers apply solely or primarily to one stage of developmental testing, indicate so by using this code (which will appear often in the questionnaire). Place the appropriate letter on the line next to your check mark(s).

- a. early stage
- b. middle stage
- c. late stage (just before field testing)

**BASED ON EXPERIENCE,
RESEARCH, OR ADVICE, WHAT
HAS PROMPTED YOU TO
UTILIZE YOUR SELECTION(S)?**

11. What do you consider the role of the developer to be during the revision of proto-type instructional materials? PLEASE CHECK, ☒ , THE MOST APPROPRIATE RESPONSE(S).

Frequency:

- 24 — ☐ the developer revises on the spot after the test subject has commented.
- 51 — ☐ the developer debriefs a small group of test subjects based on criterion test data, attitudinal survey, and elicited test subject comments. Revision is then based on these debriefing results.
- 28 — ☐ the developer revises based on data generated from teaching the instructional material to the test subject in a tutorial fashion when the subject encounters difficulties.
- 43 — ☐ the developer revises based on a comment sheet that the test subject has used to point out problem areas.
- 20 — ☐ the developer revises on the basis of criterion test data and attitudinal survey after each one-to-one session.
- 13 — ☐ OTHER (please outline briefly)
- _____
- _____
- _____
- _____

If any answers apply solely or primarily to one stage of developmental testing, indicate so by using this code (which will appear often in the questionnaire). Place the appropriate letter on the line next to it check mark(s).

a. early stage

**BASED ON EXPERIENCE,
RESEARCH, OR ADVICE, WHAT
HAS PROMPTED YOU TO
UTILIZE YOUR SELECTION(S) ?**

12. When do you stop developmental testing?
PLEASE CHECK, ☒ , THE MOST APPROPRIATE
RESPONSE(S).

Frequency:

- 6 ☐ after a pre-determined number of one-to-one sessions.
- 3 ☐ after having debriefed students from one-to-one sessions.
- 6 ☐ after having debriefed students in a small group setting.
- 26 ☐ after criterion test results begin indicating a "success" factor.
- 18 ☐ when a previously set criterion level of achievement and attitudinal response is being met during one-to-one testing.
- 29 ☐ when a previously set criterion level of achievement and attitudinal response is being met from small group debriefing sessions.
- 60 ☐ depends primarily on project deadline dates, client, money available, etc.
- 14 ☐ OTHER (please outline briefly)

**BASED ON EXPERIENCE,
RESEARCH, OR ADVICE, WHAT
HAS PROMPTED YOU TO
UTILIZE YOUR SELECTION(S) ?**

[illegible]

13. When do you stop revising? PLEASE CHECK, ☒,
THE MOST APPROPRIATE RESPONSE(S).

Frequency:

15

☐ I usually set no specified criteria for test subject performance.

2

☐ when less than 7% of the test subjects are having problems with the material.

22

☐ after 4 or 5 sessions until consistent results are obtained.

Q

☐ when there are not more than 20% unsatisfactory responses on the attitudinal survey.

0

☐ when 70% of the test subject group achieve 70% or higher on the post-test.

18

☐ when 80% of the test subject group achieve 80% or higher on the post-test.

47

☐ OTHER (please outline briefly)

**BASED ON EXPERIENCE,
RESEARCH, OR ADVICE, WHAT
HAS PROMPTED YOU TO
UTILIZE YOUR SELECTION(S) ?**

- a. early stage
- b. middle stage
- c. late stage (just before field testing)

- Frequency:**

4 26 ☐ from an attitudinal survey.

11 44 ☐ by recording (by note-taking, audio tape, or video tape) information.

32 67 ☐ by debriefing test subjects (after the
test session).

7 38 ☐ from test subject comments placed on a revision comment sheet.

10 59 from spontaneous verbal comments that the
 ☐ test subjects have made during the test
 session.

2 6 OTHER (please outline briefly) _____

**BASED ON EXPERIENCE,
RESEARCH, OR ADVICE, WHAT
HAS PROMPTED YOU TO
UTILIZE YOUR SELECTION(S)?**

15. How do you analyze the data gathered from developmental testing sessions? PLEASE CHECK, ☒, THE MOST APPROPRIATE RESPONSE(S).

Frequency:

- 45 ☐ the extent of the analysis depends on the resources available (e.g., budget and time constraints).
- 32 ☐ by comparing the cumulative effects of revisions based on a number of test sessions (e.g., by testing out the revisions generated from a previous session on the next test subject(s).
- 32 ☐ by debriefing test subjects after their completion of the material, thus providing tallied test results, attitudinal data, and test subject generated comments. The test subjects and the developer cooperatively analyze the data.
- 19 ☐ by comparing pre- and post-test data with a pre-set standard.
- 38 ☐ with strong emphasis placed on error analysis to find out why individual test subjects make errors.
- 9 ☐ OTHER (please outline briefly)

BASED ON EXPERIENCE, RESEARCH, OR ADVICE, WHAT HAS PROMPTED YOU TO UTILIZE YOUR SELECTION(S)?

Frequency:

- follow a prescribed algorithm or heuristic, e.g., certain kinds of errors dictate certain changes.

- ☐ revise on the basis of "obvious", self-evident weaknesses within the instructional material.

☐ hard to say; it is intuitive.

☐ OTHER (please outline briefly)

**BASED ON EXPERIENCE,
RESEARCH, OR ADVICE, WHAT
HAS PROMPTED YOU TO
UTILIZE YOUR SELECTION(S)?**

17. We recognize that many instructional developers have evolved sophisticated procedures over the years, for developmental testing and revision. We are interested in learning how explicit and teachable your procedures are, i.e., can they be taught to others? PLEASE CHECK, ☒ , THE MOST APPROPRIATE RESPONSE(S).

Frequency: A. With regard to developmental testing?

- 41 ☐ my procedures are well documented in the literature or in handouts, and could be/are taught to others.
- 33 ☐ my procedures could easily be documented and taught to others but have not been.
- 4 ☐ it would be difficult for my procedures to be documented or taught to others.
- 21 ☐ my procedures can not be (or are not likely to be) documented or generalized because of varying local conditions. Each project, test subject, context, etc., is different and requires a different strategy.
- 8 ☐ OTHER (please outline briefly)

**BASED ON EXPERIENCE,
RESEARCH, OR ADVICE, WHAT
HAS PROMPTED YOU TO
UTILIZE YOUR SELECTION(S) ?**

This image shows a single sheet of white paper with horizontal blue or grey ruling lines. The lines are evenly spaced and run across the width of the page. On the left side, there is a vertical margin line, creating a narrow left margin. A small portion of a binder ring or hole punch is visible at the top left corner. The paper appears to be from a standard composition notebook.

B. With regard to revision?

Frequency:

- 31 ☐ my procedures are well documented in the literature or in handouts, and could be/are taught to others.
- 33 ☐ my procedures could easily be documented and taught to others but have not been.
- 4 ☐ it would be difficult for my procedures to be documented or taught to others.
- 21 ☐ my procedures can not be (or are not likely to be) documented or generalized because of varying local conditions. Each project, test subject, context, etc., is different and requires a different strategy.
- 7 ☐ OTHER (please outline briefly)

OTHER (please outline briefly)

We would very much appreciate your enclosing or providing a reference to any documentation on developmental testing or revision that you have generated and/or use.

ENCLOSING

NOT ENCLOSING ○

**BASED ON EXPERIENCE,
RESEARCH, OR ADVICE, WHAT
HAS PROMPTED YOU TO
UTILIZE YOUR SELECTION(S)?**

Q You have described your methods of developmental testing and revision in the previous items. We recognize that everyday constraints may impose upon you procedures that you consider less than ideal. Would you take a couple of minutes to describe below what you consider to be the ideal procedures for developmental testing and revision, considering, if possible, the variables noted in the questionnaire (e.g., number of test subjects, role of the developer, etc.?)

Q



****THERE IS NO NEED TO COMPLETE THIS ITEM IF YOU HAVE FILLED OUT THE REST OF THE QUESTIONNAIRE****

19. As indicated in the cover letter, you are filling out this page because you have a different method for developing and improving instructional material than the one we described (i.e., developmental testing and revision). Would you kindly take a minute or two to describe briefly how you do it?



20. Please just check the box if you wish to be kept up to date on the research project.

I would like to receive:

- ☐ AERA papers
- ☐ Results of this study
- ☐ Both

Dear Colleague:

In early July we sent you a survey-questionnaire about developmental testing and revision. As of August 20th, we have not received your responses and are anxious to include your responses and comments in our data base. The self-addressed envelope included in the first mailing should facilitate your response. Thanks for your cooperation.

George G. Gies
Cynthia B. Weston
Charles W. Bunt

APPENDIX D

Instructions:

Please answer the following questions. The information will be kept strictly confidential.

1. Mother tongue:
2. Age (years):
3. The number of years of schooling (not including the present CEGEP year):
4. Please list your previous courses in biology and chemistry (including the present CEGEP year):

OVERVIEW

As a nurse , it is very important for you to maintain the chain of asepsis during all clinical procedures. . You will learn the specific skills needed to keep the dental environment and instruments clean and sterile in the module Sterilization and Disinfection. This submodule will provide you with background information and knowledge of microorganisms and their relationship to the sterilization and disinfection process.

INTRODUCTION

Microbiology studies microorganisms, or microscopic forms of life. These microscopic organisms can be classified or separated into four distinct groups.

OBJECTIVES

1. Define and spell terminology
 - a. microbiology
 - b. bacteria
 - c. viruses
 - d. fungi
 - e. protozoa
 - f. cocci
 - g. bacilli
 - h. spirochetes
 - i. microscope
 - j. cultures
 2. Name and describe the four different classifications of microorganisms.
 3. Describe the techniques scientists use to study microorganisms.
 4. Name and describe the three general shapes of bacteria.
 5. Differentiate between the four types of microorganisms.
-

MICROBIOLOGY
DEFINED

Microbiology is the name given to the branch science that studies forms of life which are microscopic, or not visible to the naked eye. These microscopic forms of life are called microorganisms. Microbiology studies how microorganisms live and die, and the helpful and harmful effects microorganisms have on other living things, including man.

CLASSIFICATION
OF MICRO-
ORGANISMS

One useful way of studying microscopic forms of life is to group them into general classifications. Four general classifications will be discussed here:

1. Bacteria
 2. Viruses
 3. Fungi
 4. Protozoa
-

BACTERIA

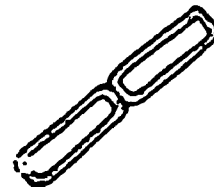
Bacteria are very small microscopic plants that are unicellular, which means that they are composed of only one cell. Bacteria have a rigid cell wall, are usually

transparent, and are no larger than 1/50,000 of an inch in size. If bacteria were large enough to see with the naked eye you would find that there are three different shapes.

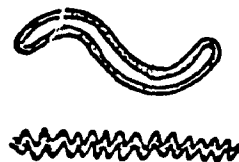
1. spheres called cocci:



2. rods called bacilli:



3. spirals called spirochetes, vibrios, or spirillum:



VIRUSES

Viruses are submicroscopic, or even smaller than bacteria. Viruses exist in several shapes and sizes, but the viruses effecting man and animals are usually spherical in shape.

FUNGI

Fungi are usually larger than either viruses or bacteria, but many are still microscopic. Fungi are plant-like microorganisms such as molds and yeasts.

PROTOZOA

Protozoa are the largest of the four types of microorganisms. They are usually unicellular and are sometimes large enough to see with the naked eye.

THE MICROSCOPE

Most organisms studied in microbiology are so small that they cannot be seen by the naked eye. Therefore, the scientist must use special instruments and special techniques to study microorganisms. One instrument used is the microscope. The microscope is a precision instrument used to magnify or enlarge the microorganism to many times its normal size so that it can be readily studied. Magnification by the microscope can be as little as two times or as much as 3,000 times the normal size of the microorganism.

STAINING

Frequently the microorganism must be prepared in a special way so that it may be viewed with the microscope. Sometimes it is placed in water so that it can be viewed alive. At other times, the microorganism or parts of its structure are difficult to see, so it is stained with a special solution to make it more clearly visible.

For example, Gram stain is a commonly used staining technique. It differentiates between bacteria by dividing them into two groups, Gram Positive or Gram Negative.

CULTURING

Another technique used by scientists to study microorganisms is known as culturing. In this technique, scientists in the laboratory observe the growth of microorganisms on specially prepared plates, dishes, or in test tubes. A special food known as a culture medium is placed in the dishes. The microorganisms are then placed in the culture medium and allowed to grow. By using different culture media and exposing the dishes to different environmental factors, the scientists learn about the microorganisms.

BIOCHEMICAL TESTS

Another way to study microorganisms is through the use of different biochemical tests. Such tests demonstrate the presence of different systems within the cell. One such test is fermentation, in which the microorganism changes sugar into alcohol.

ANIMAL
INOCULATION

Another technique used to study microorganisms is to introduce them into laboratory animals such as mice or rats and observe the animals for changes. This is known as animal inoculation. Sometimes the animal is killed and examined for the presence of internal reactions to the microorganism.

1. What are the four general classifications of microorganisms?

a. _____

b. _____

c. _____

d. _____

2. Spherical-shaped bacteria are called

- a. bacilli
- b. spirochetes
- c. vibrios
- d. cocci

3. Which of the four classifications contain microorganisms sometimes large enough to see with the naked eye?

4. A precision instrument used by scientists to magnify microorganisms for study is known as the

5. Fermentation is an example of

- a. staining
 - b. a biochemical test
 - c. culturing
 - d. inoculation
-

ATTITUDINAL QUESTIONNAIRE

NAME: _____ DATE: _____

LESSON TITLE: Sterilization and Disinfection 1

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

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. The way the material was formatted often distracted my attention.

1 2 3 4 5

5. Going through the material became tedious, or boring.

1 2 3 4 5

6. This lesson was very well organized. The concepts were highly related to each other.

T 2 3 4 5

7. There was too much information being conveyed in the lesson.

T 2 3 4 5

8. There was too much redundancy. I was bored by the repetition of ideas.

T 2 3 4 5

9. There was a lot of irrelevant information in this lesson.

T 2 3 4 5

10. The lesson was excellently designed. I could easily follow the instructions.

T 2 3 4 5

11. This lesson had very serious gaps and lacked internal continuity.

T 2 3 4 5

12. The examples used to illustrate main points were excellent.

T 2 3 4 5

13. The vocabulary used contained many unfamiliar words. I often did not understand what was going on.

T 2 3 4 5

14. The Self-Test at the end of the lesson did a good job of testing my knowledge of the main points in the lesson.

T 2 3 4 5

15. I had no idea of how I was doing during the lesson.

T 2 3 4 5

16. At the end of the lesson I was still uncertain about a lot of things and had to guess on many of the questions on the Self-Test.

T 2 3 4 5

17. I believe I learned a lot, considering the time spent on this lesson.

T 2 3 4 5

18. I would recommend extensive modifications to the lesson before using it with other students.

T 2 3 4 5



19. For you, what was the most difficult part of the lesson? _____

20. What was the easiest part of the lesson? _____

21. What were the three worst things about this lesson? _____

22. I understood most of the concepts and vocabulary immediately after completing the lesson.

1 2 3 4 5

23. I think this whole procedure of trying out new materials with students is a waste of time.

1 2 3 4 5

24. I would prefer a textbook or lecture version of this lesson rather than the workbook version.

1 2 3 4 5

24. I often needed to go back over a portion of the lesson to fully understand it.

1 2 3 4 5

25. After completing the lesson, I was more interested in and/or favorably impressed with the general subject matter than I was before the lesson.

1 2 3 4 5



26. Please write below any comments, suggestions or changes which you believe will improve this lesson. Thank you.

MICROBIOLOGY RELATED TO STERILIZATION AND DISINFECTION Page 11
ANSWERS TO SELF-TEST 1

1. a. bacteria
 b. viruses
 c. fungi
 d. protozoa
 2. d
 3. protozoa
 4. microscope
 5. b
-

APPENDIX E

INTRODUCTION

Certain conditions in the environment promote growth of microorganisms while other conditions actually slow growth or kill microorganisms. Depending upon these conditions, the microorganisms can produce disease in humans or they can be rejected by barriers produced by the body.

OBJECTIVES

1. Define and spell terminology:

a. aerobic	g. cross infection
b. anaerobic	h. virulent
c. resident microorganism	i. immunity--natural, artificial, active-passive
d. infection	j. phagocytosis
e. autogenic infection	k. portal of entry
f. environment	l. immunity--natural

2. List at least four types of environmental factors which affect microorganisms.
 3. Describe the role microorganisms play in the process of infection.
 4. List and explain the barrier the body has to infection.
 5. List the different types of immunity and give an example of each.
 6. List the points (portals) of entry of microorganisms.
 7. Describe and give an example of each method of disease transmission.
-

ENVIRONMENTAL CONDITIONS

The environmental conditions which affect the growth and multiplication of microorganisms include:

1. type of food available;
2. moisture
3. temperature
4. oxygen requirements
5. amount of light present

It is important to realize that different microorganisms require varying environmental conditions for optimal growth. These

conditions become especially important in relation to sterilization and disinfection.

RELATIONSHIP TO
STERILIZATION
AND DISINFECTION

It is essential to know which of these environmental conditions can be controlled to destroy harmful microorganisms and attain sterilization. For example, if we know a certain microorganism requires a temperature of 98 to 100 degrees Fahrenheit to live, then raising the temperature several degrees above that level may kill it. This is essentially what is done during the sterilization process.

A program to control harmful microorganisms in the hospital environment consists of those sterilization methods which kill the most resistant and harmful microorganisms.

FOOD, MOISTURE,
AND TEMPERATURE

Microorganisms each have their own specific requirements for these three environmental conditions. Much research has been conducted to determine the combinations of these conditions necessary to maintain the life of harmful microorganisms. Sterilization often changes one or all of these conditions to kill the microorganism and thus make it harmless.

OXYGEN AND
LIGHT

Oxygen

Microorganisms also differ in their need for oxygen. Aerobic microorganisms require oxygen in their environment in order to grow and multiply. Anaerobic microorganisms require an absence of oxygen in their environment in order to grow. In other words, if any oxygen is present in the environment, the anaerobic microorganisms will die.

Light

Microorganisms also differ in the amount of light they require to grow and multiply. Some require a great deal of light while others actually require darkness to promote growth.

VIRUSES

Viruses have varying requirements for environmental conditions in which to grow and multiply. Specific viruses require specific conditions. One thing that is essential to the growth of any virus is the presence of some living tissue.

MICROBIOLOGY RELATED TO STERILIZATION AND DISINFECTION Page 15
ROLE OF MICROORGANISMS IN PRODUCING DISEASE

**MICROORGANISMS
IN THE BODY**

Some microorganisms are helpful to man and some are harmful. It is possible that in some quantities some microorganisms have no effect on the body, but in larger quantities may produce disease. The body varies in its response to the presence of microorganisms.

**RESIDENT
MICROORGANISMS**

Some microorganisms are normally present within the body. These are known as resident microorganisms. The oral cavity is a good example of a part of the body where resident microorganisms may be found. It is important to understand that the resident microorganisms normally found in one person's mouth may be different from those microorganisms found in someone else's mouth. The resident microorganisms found in the mouth may not be normally present in other parts of the body.

**TRANSIENT
MICROORGANISMS**

Transient microorganisms are not normally present in the body. They are transient, or present temporarily, in the body. These microorganisms may be present in the environment or be the resident microorganisms from someone else's body. In addition, microorganisms which are resident to one area of the body may be transient to another area of the body.

For example, certain microorganisms are normally found on the finger or are resident there. If you rub your eye with your finger, some of the resident microorganisms may be transferred to the eye, where they are not normally found. They would then be called transient microorganisms and in some instances could cause disease.

**TYPES OF
INFECTION**

Infection: When a microorganism enters the body, multiplies, and causes a reaction, this process is known as an infection.

Autogenous infection: An infection is known as an autogenous infection when it is caused by a resident microorganism. It is usually the result of resident microorganisms penetrating into an area where they do not normally reside. For example, if resident

microorganisms in the mouth penetrate the oral tissues during an injection of local anesthesia and cause an infection, this would be known as an autogenous infection.

Cross infection: An infection is known as a cross infection when it is caused by microorganisms transferred from one person to another. For example, if you did not wash your hands properly (or wear gloves) and a microorganism from your hand entered a patient's mouth and caused an infection, it would be known as a cross infection.

Carrier of infection: A "carrier" is a person who carries or has present in the body a disease-producing microorganism, but does not exhibit observable symptoms of the disease. However, the carrier can transmit the disease-producing microorganism and the disease to any other individual who may be susceptible to the particular microorganism. For example, an individual may carry the microorganism that causes hepatitis and yet never exhibit the symptoms of the disease or every have hepatitis. When the carrier comes into contact with others (s)he can transmit the actual disease to the other individuals.

FACTORS NECESSARY FOR THE DEVELOPMENT OF INFECTION

Several factors are necessary for the development of an infection.

1. The microorganism must enter the body in sufficient numbers.
 2. The microorganism must be virulent. This means that it must be capable of overcoming the body's defenses and capable of destroying healthy tissue.
 3. The body must be susceptible to the disease. In other words, if the body has developed immunity to the disease, it will not be affected by the microorganism.
 4. The microorganism must be transmitted through the proper route, or portal of entry.
-

MICROBIOLOGY RELATED TO STERILIZATION AND DISINFECTION Page 17
ROLE OF MICROORGANISMS IN PRODUCING DISEASE

PORTALS OF ENTRY For infection to occur, it is necessary for the microorganism to enter the body via an appropriate route. These points, or portals of entry, include:

1. the skin, usually at the site of a cut or scrape;
2. the respiratory tract;
3. the oral cavity;
4. the nasal passages;
5. the digestive tract;
6. the genitourinary system;
7. the placenta;
8. any site where trauma occurs.

A microorganism at one site may not cause a problem, but may cause an infection when it enters at the appropriate portal of entry. For example, a microorganism which causes a respiratory infection must enter through the mouth or nose to produce an infection in the lungs.

**FUNCTION OF
BARRIERS TO
INFECTION**

In order to produce a disease or infection, the microorganism must overcome various barriers on the body's surface or at the portals of entry. The function of the barriers, then, is to prevent the microorganisms from entering the body and therefore prevent disease.

**CATEGORIES OF
BARRIERS**

The body has a number of different barriers which prevent microorganisms from producing disease. These barriers may be categorized as 1) mechanical; 2) chemical; and 3) physiologic.

**MECHANICAL
BARRIERS**

The human body is designed in such a way that it actively and mechanically prevents a great number of microorganisms from entering it. The skin is a good example of such a mechanical barrier. The cells in the top layer of skin are "tough" and prevent microorganisms from entering the body, much like the skin of an orange protects the fruit inside. The body also has various reflexes which provide mechanical barriers to microorganisms. Examples of protective

reflexes are the blink of an eye or a sneeze. Both of these reflexes help to prevent microorganisms from entering the body.

CHEMICAL BARRIERS The second category of barriers are chemical in nature. Various body fluids contain ingredients which destroy microorganisms. An example of this type of barrier is the fluid known as gastric juice, found in the stomach. The gastric juices contain certain chemicals which help to destroy microorganisms and thus resist infection.

**PHYSIOLOGIC
BARRIERS**

Physiology is the science that deals with the function of various parts and organs of living organisms.

Physiologic barriers, the largest category of barriers to infection, are the natural barriers produced by normal body function. These body functions resist the invasion of microorganisms. Three examples of physiologic barriers are phagocytosis, temperature increase, and immunity.

**PHYSIOLOGIC
BARRIERS:
PHAGOCYTOSIS**

Phagocytosis is a natural defense mechanism. In this process, certain white blood cells seek out disease-producing microorganisms, surround them, and actually destroy them.

**PHYSIOLOGIC
BARRIERS:
TEMPERATURE
INCREASE**

The body also reacts to microorganisms by an increase in temperature. The proper temperature is needed by microorganisms in order to grow and multiply. When the body's temperature rises, microorganisms are no longer able to grow and are usually destroyed.

**PHYSIOLOGIC
BARRIERS:
IMMUNITY**

The body's most complex barrier to microorganisms is known as immunity. Immunity is a complex reaction that resists invasion and disease production ability of microorganisms. Immunity may be broken down into the following categories:

MICROBIOLOGY RELATED TO STERILIZATION AND DISINFECTION Page 19
ROLE OF MICROORGANISMS IN PRODUCING DISEASE

I. Natural or inherent immunity. Natural immunity occurs in certain races and individuals.

A. Race. Some races of people are resistant to certain diseases and the microorganisms that produce those diseases.

B. Individual. Some individuals have something in their genetic structure that makes the individual naturally resistant to certain diseases.

II. Acquired immunity. Immunity acquired after conception can be natural or artificial in nature.

A. Natural.

1. Active. This immunity is acquired by the attack of a disease.

2. Passive. This immunity is transferred from mother to child through the placenta during pregnancy.

B. Artificial.

1. Active. Vaccination produces this immunity. By introducing specific microorganisms into the body (in small amounts) the body's immune system is activated.

2. Passive. By injecting a specific substance into the body, immunity is produced for a short period of time.

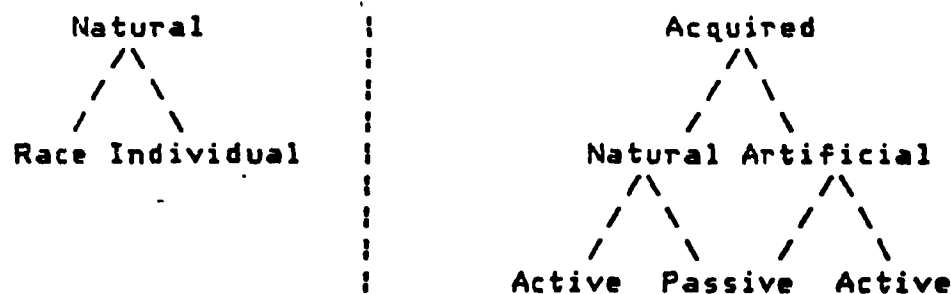


Fig. 1. Diagram of Immunity Systems.

MICROBIOLOGY RELATED TO STERILIZATION AND DISINFECTION Page 21
SELF-TEST 2

1. Name at least three of the five environmental conditions which affect the growth of microorganisms.

A. _____

B. _____

C. _____

D. _____

E. _____
2. Microorganisms which require oxygen in their environment are called _____.
3. What special environmental conditions do viruses require that other microorganisms do not require?

4. Microorganisms normally present in the body are known as _____ microorganisms, while those which are temporary are known as _____ microorganisms.
5. When a microorganism enters the body, multiplies and causes a reaction, the process is known as a(n) _____.

MICROBIOLOGY RELATED TO STERILIZATION AND DISINFECTION Page 22
SELF-TEST 2

6. An infection caused by resident microorganisms is known as a(n)

- a. cross infection
- b. autogenous infection
- c. transient infection
- d. carrier infection

7. What four factors are necessary for the development of infection?

- A. _____
- B. _____
- C. _____
- D. _____

8. Name at least five of the eight portals of entry discussed in the module.

A. _____	E. _____
B. _____	F. _____
C. _____	G. _____
D. _____	H. _____

9. The skin is an example of a(n) _____ barrier to infection.

10. Which of the following is not considered to be a physiologic barrier?
- a. gastric juice
 - b. phagocytosis
 - c. immunity
 - d. body temperature
11. Individuals who have a resistance to disease because of their genetic structure have what type of immunity?
- _____
12. Vaccinations are considered to be what type of immunity?
- _____
13. Three methods of disease transmission include insect carrier, _____, and _____.
14. Airborne transmission of microorganisms would be an _____ example of _____ the _____ method of disease transmission.
- _____

1. In the past, have you ever been exposed to subject matter content related to the topics of sterilization and disinfection in Microbiology?

Please check ☐ one of the following boxes.

☐ - NO

☐ - YES - If yes, please answer #2.

2. I have previously been exposed to similar subject matter content in microbiology through:

Please check the appropriate box(es):

☐ a course in Microbiology

☐ reading books

☐ work or work-related activities

ATTITUDINAL QUESTIONNAIRE

NAME: _____ DATE: _____

LESSON TITLE: Sterilization and Disinfection 2

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

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. The way the material was formatted often distracted my attention.

1 2 3 4 5

5. Going through the material became tedious, or boring.

1 2 3 4 5

6. This lesson was very well organized. The concepts were highly related to each other.

T 2 3 4 5

7. There was too much information being conveyed in the lesson.

T 2 3 4 5

8. There was too much redundancy. I was bored by the repetition of ideas.

T 2 3 4 5

9. There was a lot of irrelevant information in this lesson.

T 2 3 4 5

10. The lesson was excellently designed. I could easily follow the instructions.

T 2 3 4 5

11. This lesson had very serious gaps and lacked internal continuity.

T 2 3 4 5

12. The examples used to illustrate main points were excellent.

T 2 3 4 5

13. The vocabulary used contained many unfamiliar words. I often did not understand what was going on.

T 2 3 4 5

14. The Self-Test at the end of the lesson did a good job of testing my knowledge of the main points in the lesson.

T 2 3 4 5

15. I had no idea of how I was doing during the lesson.

T 2 3 4 5

16. At the end of the lesson I was still uncertain about a lot of things and had to guess on many of the questions on the Self-Test.

T 2 3 4 5

17. I believe I learned a lot, considering the time spent on this lesson.

T 2 3 4 5

18. I would recommend extensive modifications to the lesson before using it with other students.

T 2 3 4 5

0

19. For you, what was the most difficult part of the lesson? _____

20. What was the easiest part of the lesson? _____

21. What were the three worst things about this lesson? _____

22. I understood most of the concepts and vocabulary immediately after completing the lesson.

1 2 3 4 5

23. I think this whole procedure of trying out new materials with students is a waste of time.

1 2 3 4 5

24. I would prefer a textbook or lecture version of this lesson rather than the workbook version.

1 2 3 4 5

24. I often needed to go back over a portion of the lesson to fully understand it.

1 2 3 4 5

25. After completing the lesson, I was more interested in and/or favorably impressed with the general subject matter than I was before the lesson.

1 2 3 4 5

0

26. Please write below any comments, suggestions or changes which you believe will improve this lesson. Thank you.

1. A. food
 B. moisture
 C. temperature
 D. oxygen
 E. light

 See page 12-13

2. aerobic; see page 13.

3. presence of living tissue; see page 14.

4. resident, transient; see page 15.

5. infection; see page 15.

6. b; see page 15.

7. A. sufficient numbers of microorganisms
 B. virulent microorganisms
 C. susceptible person
 D. proper portal of entry of microorganisms

 See page 16.

8. A. skin E. digestive tract
 B. respiratory tract F. genitourinary system
 C. oral cavity G. placenta
 D. nasal passages H. site of trauma

 See page 17.

9. mechanical; see page 17.

MICROBIOLOGY RELATED TO STERILIZATION AND DISINFECTION Page 25
ANSWERS TO SELF-TEST 2

- 10. a; see page 18.
 - 11. natural or inherent; see page 19.
 - 12. acquired, artificial, active; see page 19.
 - 13. direct contact, indirect contact; see page 20.
 - 14. indirect contact; see page 20.
-

APPENDIX F

Consent form for learners

I have voluntarily agreed to participate in this study.

I understand that it is an investigation of various procedures for revising and improving instructional materials during their development.

I understand that my task is to act as a learner using the instructional materials presented to me.

I understand that my identity will be protected and that all records will be coded for anonymity.

Signed: _____

Dated: _____

APPENDIX 8

Experimental Condition: One-to-One Active

INSTRUCTIONS

The following learning materials are in first draft form. The purpose of this session is NOT to test how well you can learn, but how well the materials teach. Your feedback on the effectiveness of the material is important. Go through the material as best you can and if either the instructor or yourself notices a problem with the material, it will be discussed during the session. Thank you very much for your cooperation.

Experimental Condition: Small group Active

INSTRUCTIONS

The following learning materials are in first draft form. The purpose of this session is NOT to test how well you can learn, but how well the materials teach. Your feedback on the effectiveness of the material is important. Go through the material as best you can and if either the instructor or yourself notices a problem with the material, it will be discussed during the session. Please do not talk with your "neighbors" during the session. Thank you very much for your cooperation.

Experimental Condition: One-to-One Passive

INSTRUCTIONS

The following learning materials are in first draft form. The purpose of this session is NOT to test how well you can learn, but how well the materials teach. Your feedback on the effectiveness of the material is important. Go through the material as best you can and if you encounter difficulty within the material, please underline the section(s) of the text, diagrams, etc., that you feel are problematic. After underlining the problem area(s), write on the materials themselves what you think is the reason for the difficulty you have encountered. Finally, if you feel that you can, suggest a way to revise the material so that another student would not encounter the same problem (please write out your revision on the material itself). If you can't suggest a revision, don't worry; just move on within the material. Thank you very much for your cooperation.

INSTRUCTIONS

The following learning materials are in first draft form. The purpose of this session is NOT to test how well you can learn, but how well the materials teach. Your feedback on the effectiveness of the material is important. Go through the material as best you can and if you encounter difficulty within the material, please underline the section(s) of the text, diagrams, etc., that you feel are problematic. After underlining the problem area(s), write on the materials themselves what you think is the reason for the difficulty you have encountered. Finally, if you feel that you can, suggest a way to revise the material so that another student would not encounter the same problem (please write out your revision on the material itself). If you can't suggest a revision, don't worry; just move on within the material. Please do not talk with your "neighbors" during the session. Thank you very much for your cooperation.

APPENDIX H

O

POSTTEST - STERILIZATION & DISINFECTION 2
SUMMARY OF RESULTS

Items (n=22)	1a	1b	1c	2	3	4a	4b	5	6	7a	7b	7c	7d	8a	8b	8c	8d	8e	9	10	11	12	%
1. name																							
2.																							
3.																							
4.																							
5.																							
6.																							
Total students correct																							
Percent students correct																							
Agenda																							

NOTE THAT AN ITEM GOES ON THE AGENDA IF 2 OR MORE STUDENTS GET THE ITEM WRONG

O

[illegible]

[illegible]

ITEM •

16.

17.

18.

19.

20.

21.

22.

23.

24.

25.

26.

APPENDIX I

Table 11.

Descriptive Statistics for the Number of Problems Identified Unit #1 for the Total Sample of Test Subjects

<u>Number of Problems Identified</u>	<u>Frequency</u>	<u>Percent</u>
0	39	54.2
1	12	16.7
2	8	11.1
3	2	2.8
4	1	1.4
5	3	4.2
6	3	4.2
9	1	1.4
11	1	1.4
12	1	1.4
16	1	1.4
Total	72	100.0

Mean = 1.65

SD = 3.06

Table 12.

Descriptive Statistics for the Number of Problems Identified Unit #2 for the Total Sample of Test Subjects

<u>Number of Problems Identified</u>	<u>Frequency</u>	<u>Percent</u>
0	31	43.7
1	9	12.5
2	5	6.9
3	8	11.1
4	5	6.9
5	3	4.2
6	2	2.8
7	2	2.8
8	1	1.4
9	2	2.8
11	1	1.4
12	1	1.4
22	1	1.4
Missing	1	1.4
Total	72	100.0

Mean = 2.47

SD = 3.72

Table 13.

Descriptive Statistics for the Number of Problems Identified in Total for the Total Sample of Test Subjects

<u>Number of Problems Identified</u>	<u>Frequency</u>	<u>Percent</u>
0	26	36.1
1	11	15.3
2	2	2.8
3	4	5.6
4	7	9.7
5	4	5.6
6	1	1.4
7	1	1.4
8	2	2.8
9	5	6.9
10	2	2.8
11	1	1.4
12	1	1.4
13	2	2.8
17	1	1.4
21	1	1.4
38	1	1.4
Total	72	100.0

Mean = 4.08

SD = 6.12

Table 14.

Descriptive Statistics for the Main Effect of Role of Participants on the Number of Problems Identified
by Test Subjects for Unit #1

<u>Number of Problems Identified</u>	<u>Frequency (Active Role)</u>	<u>Percent</u>	<u>Frequency (Passive Role)</u>	<u>Percent</u>
0	21	58.3	18	50.0
1	8	22.2	4	11.1
2	4	11.1	4	11.1
3	0	0.0	2	5.6
4	0	0.0	1	2.8
5	1	2.8	2	5.6
6	0	0.0	3	8.3
9	1	2.8	0	0.0
11	0	0.0	1	2.8
12	0	0.0	1	2.8
16	1	2.8	0	0.0
Total	36	100.0	36	100.0
Mean = 1.28			Mean = 2.03	
SD = 3.06			SD = 3.07	

Table 15.

Descriptive Statistics for the Main Effect of Role of Participants on the Number of Problems Identified
by Test Subjects for Unit #2

<u>Number of Problems Identified</u>	<u>Frequency (Active Role)</u>	<u>Percent</u>	<u>Frequency (Passive Role)</u>	<u>Percent</u>
0	19	52.8	12	33.3
1	4	11.1	5	13.9
2	3	8.3	2	5.6
3	3	8.3	5	13.9
4	1	2.8	4	11.1
5	1	2.8	2	5.6
6	1	2.8	1	2.8
7	1	2.8	1	2.8
8	1	2.8	0	0.0
9	0	0.0	2	5.6
11	0	0.0	1	2.8
12	0	0.0	1	2.8
22	1	2.8	0	0.0
missing	1	2.8	0	0.0
Total	36	100.0	36	100.0
Mean = 2.03			Mean = 2.89	
SD = 4.11			SD = 3.31	

Table 16.

Descriptive Statistics for the Main Effect of Role of Participants on the Number of Problems Identified
by Test Subjects in Total

<u>Number of Problems Identified</u>	<u>Frequency (Active Role)</u>	<u>Percent</u>	<u>Frequency (Passive Role)</u>	<u>Percent</u>
0	17	47.2	9	25.0
1	5	13.9	6	16.7
2	1	2.8	1	2.8
3	1	2.8	3	8.3
4	4	11.1	3	8.3
5	2	5.6	2	5.6
6	0	0.0	1	2.8
7	1	2.8	0	0.0
8	1	2.8	1	2.8
9	2	5.6	3	8.3
10	1	2.8	1	2.8
11	0	0.0	1	2.8
12	0	0.0	1	2.8
13	0	0.0	2	5.6
17	0	0.0	1	2.8
21	0	0.0	1	2.8
38	1	2.8	0	0.0
Total	36	100.0	36	100.0
Mean = 3.25		Mean = 4.92		
SD = 6.70		SD = 5.45		

Table 17.

Descriptive Statistics for the Main Effect of Number of Participants on the Number of Problems
Identified by Test Subjects for Unit #1

<u>Number of Problems Identified Unit</u>	<u>Frequency (One-to-One)</u>	<u>Percent</u>	<u>Frequency (Small Group)</u>	<u>Percent</u>
0	17	47.2	22	61.1
1	5	13.9	7	19.4
2	5	13.9	3	8.3
3	1	2.8	1	2.8
4	0	0.0	1	2.8
5	3	8.3	0	0.0
6	1	2.8	2	5.6
9	1	2.8	0	0.0
11	1	2.8	0	0.0
12	1	2.8	0	0.0
16	1	2.8	0	0.0
Total	36	100.0	36	100.0
Mean = 2.42			Mean = 0.89	
SD = 3.92			SD = 1.58	

Table 18.

Descriptive Statistics for the Main Effect of Number of Participants on the Number of Problems
Identified by Test Subjects for Unit #2

<u>Number of Problems Identified</u>	<u>Frequency (One-to-One)</u>	<u>Percent</u>	<u>Frequency (Small Group)</u>	<u>Percent</u>
0	13	36.1	18	50.0
1	5	13.9	4	11.1
2	5	13.9	0	0.0
3	4	11.1	4	11.1
4	2	5.6	3	8.3
5	1	2.8	2	5.6
6	1	2.8	1	2.8
7	2	5.6	0	0.0
8	0	0.0	1	2.8
9	1	2.8	1	2.8
11	0	0.0	1	2.8
12	1	2.8	0	0.0
22	1	2.8	0	0.0
missing	0	0.0	1	2.8
Total	36	100.0	36	100.0
Mean = 2.86			Mean = 2.06	
SD = 4.36			SD = 2.93	

Table 19.

Descriptive Statistics for the Main Effect of Number of Participants on the Number of Problems in
Total Identified by Test Subjects

<u>Number of Problems Identified</u>	<u>Frequency (One-to-One)</u>	<u>Percent</u>	<u>Frequency (Small Group)</u>	<u>Percent</u>
0	11	30.6	15	41.7
1	5	13.9	6	16.7
2	0	0.0	2	5.6
3	3	8.3	1	2.8
4	4	11.1	3	8.3
5	3	8.3	1	2.8
6	0	0.0	1	2.8
7	1	2.8	0	0.0
8	1	2.8	1	2.8
9	2	5.6	3	8.3
10	1	2.8	1	2.8
11	0	0.0	1	2.8
12	0	0.0	1	2.8
13	2	5.6	0	0.0
17	1	2.8	0	0.0
21	1	2.8	0	0.0
22	0	0.0	0	0.0
38	1	2.8	0	0.0
Total	36	100.0	36	100.0

Mean = 5.28
SD = 7.66

Mean = 2.89
SD = 3.79

Table 20.

Descriptive Statistics for the Interactions of Role and Number of Participants on the Number of Problems Identified by Test Subjects for Unit #1

<u>Number of Problems Identified</u>	<u>Freq (#1)</u>	<u>%</u>	<u>Freq (#2)</u>	<u>%</u>	<u>Freq (#3)</u>	<u>%</u>	<u>Freq (#4)</u>	<u>%</u>
0	8	44.4	9	50.0	13	72.2	9	50.0
1	3	16.7	2	11.1	5	27.8	2	11.1
2	4	22.2	1	5.6	0	0.0	3	16.7
3	0	0.0	1	5.6	0	0.0	1	5.6
4	0	0.0	0	0.0	0	0.0	1	5.6
5	1	5.6	2	11.1	0	0.0	0	0.0
6	0	0.0	1	5.6	0	0.0	2	11.1
9	1	5.6	0	0.0	0	0.0	0	0.0
11	0	0.0	1	5.6	0	0.0	0	0.0
12	0	0.0	1	5.6	0	0.0	0	0.0
16	1	5.6	0	0.0	0	0.0	0	0.0
Total	18	100.0	18	100.0	18	100.0	18	100.0
	Mean = 2.28		Mean = 2.56		Mean = 0.28		Mean = 1.50	
	SD = 4.11		SD = 3.82		SD = 0.46		SD = 2.04	

Table 21.

Descriptive Statistics for the Interactions of Role and Number of Participants on the Number of Problems Identified by Test Subjects for Unit #2

<u>Number of Problems Identified</u>	<u>Freq (#1)</u>	<u>%</u>	<u>Freq (#2)</u>	<u>%</u>	<u>Freq (#3)</u>	<u>%</u>	<u>Freq (#4)</u>	<u>%</u>
0	6	33.3	7	38.9	13	72.2	5	27.8
1	2	11.1	3	16.7	2	11.1	2	11.1
2	3	16.7	2	11.1	0	0.0	0	0.0
3	2	11.1	2	11.1	1	5.6	3	16.7
4	1	5.6	1	5.6	0	0.0	3	16.7
5	1	5.6	0	0.0	0	0.0	2	11.1
6	1	5.6	0	0.0	0	0.0	1	5.6
7	1	5.6	1	5.6	0	0.0	0	0.0
8	0	0.0	0	0.0	1	5.6	0	0.0
9	0	0.0	1	5.6	0	0.0	1	5.6
11	0	0.0	0	0.0	0	0.0	1	5.6
12	0	0.0	1	5.6	0	0.0	0	0.0
22	1	5.6	0	0.0	0	0.0	0	0.0
missing	0	0.0	0	0.0	1	5.6	0	0.0
Total	18	100.0	18	100.0	18	100.0	18	100.0
	Mean = 3.22		Mean = 2.50		Mean = 0.77		Mean = 3.28	
	SD = 5.17		SD = 3.49		SD = 2.02		SD = 3.18	

* Key: #1=One-to-One Active; #2 =One-to-One Passive; #3 =Small Group Active; #4=Small Group Passive

Table 22.

Descriptive Statistics for the Interactions of Role and Number of Participants on the Number of Problems Identified in Total by Test Subjects

<u>Number of Problems Identified</u>	<u>Freq (#1)</u>	<u>%</u>	<u>Freq (#2)</u>	<u>%</u>	<u>Freq (#3)</u>	<u>%</u>	<u>Freq (#4)</u>	<u>%</u>
0	5	27.8	6	33.3	12	66.7	3	16.7
1	2	11.1	3	16.7	3	16.7	3	16.7
2	0	0.0	0	0.0	1	5.6	1	5.6
3	1	5.6	2	11.1	0	0.0	1	5.6
4	3	16.7	1	5.6	1	5.6	2	11.1
5	2	11.1	1	5.6	0	0.0	1	5.6
6	0	0.0	0	0.0	0	0.0	1	5.6
7	1	5.6	0	0.0	0	0.0	0	0.0
8	1	5.6	0	0.0	0	0.0	1	5.6
9	1	5.6	1	5.6	1	5.6	2	11.1
10	1	5.6	0	0.0	0	0.0	1	5.6
11	0	0.0	0	0.0	0	0.0	1	5.6
12	0	0.0	0	0.0	0	0.0	1	5.6
13	0	0.0	2	11.1	0	0.0	0	0.0
17	0	0.0	1	5.6	0	0.0	0	0.0
21	0	0.0	1	5.6	0	0.0	0	0.0
38	1	5.6	0	0.0	0	0.0	0	0.0
Total	18	100.0	18	100.0	18	100.0	18	100.0
	Mean = 5.50		Mean = 5.06		Mean = 1.00		Mean = 4.78	
	SD = 8.75		SD = 6.65		SD = 2.25		SD = 4.12	

* Key: #1 = One-to-One Active
 #2 = One-to-One Passive
 #3 = Small Group Active
 #4 = Small Group Passive

Table 23.

Means Based on the Two-Way ANOVA of the Number of Problems Identified

Role of Participants	Number of Participants	Problems Identified During DT		
		Unit #1	Unit #2	Total
—	One-to-One	2.42	2.86	5.28
—	Small Group	0.89	2.06	2.89
Active	—	1.28	2.03	3.25
Passive	—	2.03	2.89	4.92
Active	One-to-One	2.28	3.22	5.50
	Small Group	0.28	0.77	1.00
Passive	One-to-One	2.56	2.50	5.06
	Small Group	1.50	3.28	4.78

Table 25.

Descriptive Statistics for the Number of Debriefed Comments Unit #1 for the Total Sample of Test Subjects

<u>Number of Debriefed Comments</u>	<u>Frequency</u>	<u>Percent</u>
0	47	65.3
1	4	5.6
2	11	15.3
3	2	2.8
5	2	2.8
6	1	1.4
7	1	1.4
9	1	1.4
11	1	1.4
12	1	1.4
29	1	1.4
Total	72	100.0

Mean = 1.61

SD = 4.11

Table 26.

Descriptive Statistics for the Number of Debriefed Comments Unit #2 for the Total Sample of Test Subjects

<u>Number of Debriefed Comments</u>	<u>Frequency</u>	<u>Percent</u>
0	29	40.3
1	13	18.1
2	5	6.9
3	3	4.2
4	7	9.7
5	2	2.8
6	1	1.4
7	2	2.8
9	1	1.4
10	6	8.3
11	1	1.4
12	1	1.4
Missing	1	1.4
Total	72	100.0

Mean = 2.56

SD = 3.46

Table 27.

Descriptive Statistics for the Number of Debriefed Comments in Total for the Total Sample of Test Subjects

<u>Number of Debriefed Comments</u>	<u>Frequency</u>	<u>Percent</u>
0	24	33.3
1	9	12.5
2	6	8.3
3	2	2.8
4	12	16.7
5	3	4.2
7	1	1.4
8	1	1.4
9	1	1.4
10	8	11.1
13	1	1.4
14	1	1.4
18	1	1.4
23	1	1.4
38	1	1.4
Total	72	100.0

Mean = 4.17

SD = 6.19

Table 28.

Descriptive Statistics for the Main Effect of Role of Participants on the Number of Debriefed Comments for Unit #1

<u>Number of Debriefed Comments</u>	<u>Frequency (Active Role)</u>	<u>Percent</u>	<u>Frequency (Passive Role)</u>	<u>Percent</u>
0	26	72.2	21	58.3
1	2	5.6	2	5.6
2	4	11.1	7	19.4
3	1	2.8	1	2.8
5	1	2.8	1	2.8
6	0	0.0	1	2.8
7	0	0.0	1	2.8
9	1	2.8	0	0.0
11	0	0.0	1	2.8
12	0	0.0	1	2.8
29	1	2.8	0	0.0
Total	36	100.0	36	100.0
Mean = 1.56			Mean = 1.67	
SD = 5.03			SD = 2.30	

Table 29.

Descriptive Statistics for the Main Effect of Role of Participants on the Number of Debriefed Comments for Unit #2

<u>Number of Debriefed Comments</u>	<u>Frequency (Active Role)</u>	<u>Percent</u>	<u>Frequency (Passive Role)</u>	<u>Percent</u>
0	16	44.4	13	36.1
1	3	8.3	10	27.8
2	3	8.3	2	5.6
3	2	5.6	1	2.8
4	7	19.4	0	0.0
5	1	2.8	1	2.8
6	1	2.8	0	0.0
7	1	2.8	1	2.8
9	1	2.8	0	0.0
10	0	0.0	6	16.7
11	0	0.0	1	2.8
12	0	0.0	1	2.8
Missing	1	2.8	0	0.0
Total	36	100.0	36	100.0
Mean = 2.00			Mean = 3.11	
SD = 2.40			SD = 4.20	

Table 30.

Descriptive Statistics for the Main Effect of Role of Participants on the Number of Debriefed Comments in Total

<u>Number of Debriefed Comments</u>	<u>Frequency (Active Role)</u>	<u>Percent</u>	<u>Frequency (Passive Role)</u>	<u>Percent</u>
0	17	47.2	7	19.4
1	1	2.8	8	22.2
2	0	0.0	6	16.7
3	1	2.8	1	2.8
4	10	27.8	2	5.6
5	2	5.6	1	2.8
7	1	2.8	0	0.0
8	1	2.8	0	0.0
9	1	2.8	0	0.0
10	1	2.8	7	19.4
13	0	0.0	1	2.8
14	0	0.0	1	2.8
18	0	0.0	1	2.8
23	0	0.0	1	2.8
38	1	2.8	0	0.0
Total	36	100.0	36	100.0

Mean = 3.50
SD = 6.58

Mean = 4.83
SD = 5.79

Table 31.

Descriptive Statistics for the Main Effect of Number of Participants on the Number of Debriefed Comments for Unit #1

<u>Number of Debriefed Comments</u>	<u>Frequency (One-to-One)</u>	<u>Percent</u>	<u>Frequency (Small Group)</u>	<u>Percent</u>
0	17	47.2	30	83.3
1	4	11.1	0	0.0
2	5	13.9	6	16.7
3	2	5.6	0	0.0
5	2	5.6	0	0.0
6	1	2.8	0	0.0
7	1	2.8	0	0.0
9	1	2.8	0	0.0
11	1	2.8	0	0.0
12	1	2.8	0	0.0
29	1	2.8	0	0.0
Total	36	100.0	36	100.0
Mean = 2.89		Mean = 0.33		
SD = 5.51		SD = 0.76		

Table 32.

Descriptive Statistics for the Main Effect of Number of Participants on the Number of Debriefed Comments for Unit #2

<u>Number of Debriefed Comments</u>	<u>Frequency (One-to-One)</u>	<u>Percent</u>	<u>Frequency (Small Group)</u>	<u>Percent</u>
0	12	33.3	17	47.2
1	7	19.4	6	16.7
2	5	13.9	0	0.0
3	3	8.3	0	0.0
4	1	2.8	6	16.7
5	2	5.6	0	0.0
6	1	2.8	0	0.0
7	2	5.6	0	0.0
9	1	2.8	0	0.0
10	0	0.0	6	16.7
11	1	2.8	0	0.0
12	1	2.8	0	0.0
missing	0	0.0	1	2.8
Total	36	100.0	36	100.0
	Mean = 2.56		Mean = 2.57	
	SD = 3.23		SD = 3.72	

Table 33.

Descriptive Statistics for the Main Effect of Number of Participants on the Number of Debriefed
Comments in Total

<u>Number of Debriefed Comments</u>	<u>Frequency (One-to-One)</u>	<u>Percent</u>	<u>Frequency (Small Group)</u>	<u>Percent</u>
0	12	33.3	12	33.3
1	3	8.3	6	16.7
2	0	0.0	6	16.7
3	2	5.6	0	0.0
4	6	16.7	6	16.7
5	3	8.3	0	0.0
7	1	2.8	0	0.0
8	1	2.8	0	0.0
9	1	2.8	0	0.0
10	2	5.6	6	16.7
13	1	2.8	0	0.0
14	1	2.8	0	0.0
18	1	2.8	0	0.0
23	1	2.8	0	0.0
38	1	2.8	0	0.0
Total	36	100.0	36	100.0

Mean = 5.50
SD = 7.85

Mean = 2.83
SD = 3.53

Table 34.

Descriptive Statistics for the Interactions of Role and Number of Participants on the Number of Debriefed Comments for Unit #1

<u>Number of Debriefed Comments</u>	<u>Freq (#1)</u>	<u>%</u>	<u>Freq (#2)</u>	<u>%</u>	<u>Freq (#3)</u>	<u>%</u>	<u>Freq (#4)</u>	<u>%</u>
0	8	44.4	9	50.0	18	100.0	12	66.7
1	2	11.1	2	11.1	0	0.0	0	0.0
2	4	22.2	1	5.6	0	0.0	6	33.3
3	1	5.6	1	5.6	0	0.0	0	0.0
5	1	5.6	1	5.6	0	0.0	0	0.0
6	0	0.0	1	5.6	0	0.0	0	0.0
7	0	0.0	1	5.6	0	0.0	0	0.0
9	1	5.6	0	0.0	0	0.0	0	0.0
11	0	0.0	1	5.6	0	0.0	0	0.0
12	0	0.0	1	5.6	0	0.0	0	0.0
29	1	5.6	0	0.0	0	0.0	0	0.0
Total	18	100.0	18	100.0	18	100.0	18	100.0
	Mean = 3.11		Mean = 2.67		Mean = 0.00		Mean = 0.67	
	SD = 6.86		SD = 3.93		SD = 0.00		SD = 0.97	

Table 35.

Descriptive Statistics for the Interactions of Role and Number of Participants on the Number of Debriefed Comments for Unit #2

<u>Number of Debriefed Comments</u>	<u>Freq (#1)</u>	<u>%</u>	<u>Freq (#2)</u>	<u>%</u>	<u>Freq (#3)</u>	<u>%</u>	<u>Freq (#4)</u>	<u>%</u>
0	5	27.8	7	38.9	11	61.1	6	33.3
1	3	16.7	4	22.2	0	0.0	6	33.3
2	3	16.7	2	11.1	0	0.0	0	0.0
3	2	11.1	1	5.6	0	0.0	0	0.0
4	1	5.6	0	0.0	6	33.3	0	0.0
5	1	5.6	1	5.6	0	0.0	0	0.0
6	1	5.6	0	0.0	0	0.0	0	0.0
7	1	5.6	1	5.6	0	0.0	0	0.0
9	1	5.6	0	0.0	0	0.0	0	0.0
10	0	0.0	0	0.0	0	0.0	6	33.3
11	0	0.0	1	5.6	0	0.0	0	0.0
12	0	0.0	1	5.6	0	0.0	0	0.0
missing	0	0.0	0	0.0	1	5.6	0	0.0
Total	18	100.0	18	100.0	18	100.0	18	100.0
	Mean = 2.56		Mean = 2.56		Mean = 1.41		Mean = 3.67	
	SD = 2.68		SD = 3.78		SD = 1.97		SD = 4.63	

* Key: #1=One-to-One Active; #2=One-to-One Passive; #3=Small Group Active; #4=Small Group Passive

Table 36.

Descriptive Statistics for the Interactions of Role and Number of Participants on the Number of Debriefed Comments in Total

<u>Number of Debriefed Comments</u>	<u>Freq (#1)</u>	<u>%</u>	<u>Freq (#2)</u>	<u>%</u>	<u>Freq (#3)</u>	<u>%</u>	<u>Freq (#4)</u>	<u>%</u>
0	5	27.8	7	38.9	12	66.7	0	0.0
1	1	5.6	2	11.1	0	0.0	6	33.3
2	0	0.0	0	0.0	0	0.0	6	33.3
3	1	5.6	1	5.6	0	0.0	0	0.0
4	4	22.2	2	11.1	6	33.3	0	0.0
5	2	11.1	1	5.6	0	0.0	0	0.0
7	1	5.6	0	0.0	0	0.0	0	0.0
8	1	5.6	0	0.0	0	0.0	0	0.0
9	1	5.6	0	0.0	0	0.0	0	0.0
10	1	5.6	1	5.6	0	0.0	6	33.3
13	0	0.0	1	5.6	0	0.0	0	0.0
14	0	0.0	1	5.6	0	0.0	0	0.0
18	0	0.0	1	5.6	0	0.0	0	0.0
23	0	0.0	1	5.6	0	0.0	0	0.0
38	1	5.6	0	0.0	0	0.0	0	0.0
Total	18	100.0	18	100.0	18	100.0	18	100.0
	Mean = 5.67		Mean = 5.33		Mean = 1.33		Mean = 4.33	
	SD = 8.69		SD = 7.17		SD = 1.94		SD = 4.14	

* Key: #1 = One-to-One Active
 #2 = One-to-One Passive
 #3 = Small Group Active
 #4 = Small Group Passive

Table 37.

Means Based on the Two-Way ANOVA of the Number of Debriefed Comments

Role of Participants	Number of Participants	Debriefed Comments		
		Unit #1	Unit #2	Total
—	One-to-One	2.89	2.56	5.50
—	Small Group	0.33	2.57	2.83
Active	—	1.56	2.00	3.50
Passive	—	1.67	3.11	4.83
Active	One-to-One	3.11	2.56	5.67
	Small Group	0.00	1.41	1.33
Passive	One-to-One	2.67	2.56	5.33
	Small Group	0.67	3.67	4.33

Table 39.

Descriptive Statistics for the Number of Debriefed Posttest Errors Unit #1 by the Total Sample of Test Subjects

<u>Number of Debriefed Posttest Errors</u>	<u>Frequency</u>	<u>Percent</u>
0	15	20.8
1	24	33.3
2	11	15.3
3	21	29.2
4	1	1.4
Total	72	100.0

Mean = 1.57

SD = 1.16

Table 40.

Descriptive Statistics for the Number of Debriefed Posttest Errors Unit #2 by the Total Sample of Test Subjects

<u>Number of Debriefed Posttest Errors</u>	<u>Frequency</u>	<u>Percent</u>
0	1	1.4
1	7	9.7
2	9	12.5
3	11	15.3
4	8	11.1
5	25	34.7
6	9	12.5
7	1	1.4
Missing	1	1.4
Total	72	100.0

Mean = 3.89

SD = 1.67

Table 41.

Descriptive Statistics for the Number of Debriefed Posttest Errors in Total by the Total Sample of Test Subjects

<u>Number of Debriefed Posttest Errors</u>	<u>Frequency</u>	<u>Percent</u>
0	1	1.4
2	14	19.4
3	4	5.6
4	9	12.5
5	3	4.2
6	17	23.6
7	4	5.6
8	13	18.1
9	6	8.3
10	1	1.4
Total	72	100.0

Mean = 5.40

SD = 2.48

Table 42.

Descriptive Statistics for the Main Effect of Role of Participants on the Number of Debriefed Posttest Errors for Unit #1

<u>Number of Debriefed Posttest Errors</u>	<u>Frequency (Active Role)</u>	<u>Percent</u>	<u>Frequency (Passive Role)</u>	<u>Percent</u>
0	10	27.8	5	13.9
1	13	36.1	11	30.6
2	4	11.1	7	19.4
3	8	22.2	13	36.1
4	1	2.8	0	0.0
Total	36	100.0	36	100.0

Mean = 1.36
SD = 1.20

Mean = 1.78
SD = 1.10

Table 43.

Descriptive Statistics for the Main Effect of Role of Participants on the Number of Debriefed Posttest Errors for Unit #2

<u>Number of Debriefed Posttest Errors</u>	<u>Frequency (Active Role)</u>	<u>Percent</u>	<u>Frequency (Passive Role)</u>	<u>Percent</u>
0	1	2.8	0	0.0
1	1	2.8	6	16.7
2	7	19.4	2	5.6
3	6	16.7	5	13.9
4	2	5.6	6	16.7
5	16	44.4	9	25.0
6	2	5.6	7	19.4
7	0	0.0	1	2.8
8	0	0.0	0	0.0
9	0	0.0	0	0.0
11	0	0.0	0	0.0
12	0	0.0	0	0.0
22	0	0.0	0	0.0
missing	1	2.8	0	0.0
Total	36	100.0	36	100.0

Mean = 3.80
SD = 1.55

Mean = 3.97
SD = 1.80

Table 44.

Descriptive Statistics for the Main Effect of Role of Participants on the Number of Debriefed Posttest Errors in Total

<u>Number of Debriefed Posttest Errors</u>	<u>Frequency (Active Role)</u>	<u>Percent</u>	<u>Frequency (Passive Role)</u>	<u>Percent</u>
0	1	2.8	0	0.0
2	7	19.4	7	19.4
3	2	5.6	2	5.6
4	5	13.9	4	11.1
5	2	5.6	1	2.8
6	8	22.2	9	25.0
7	4	11.1	0	0.0
8	7	19.4	6	16.7
9	0	0.0	1	2.8
10	0	0.0	0	0.0
Total	36	100.0	36	100.0

Mean = 5.06
SD = 2.32

Mean = 5.75
SD = 2.61

Table 45.

Descriptive Statistics for the Main Effect of Number of Participants on the Number of Debriefed Posttest Errors for Unit #1

<u>Number of Debriefed Posttest Errors</u>	<u>Frequency (One-to-One)</u>	<u>Percent</u>	<u>Frequency (Small Group)</u>	<u>Percent</u>
0	10	27.8	5	13.9
1	18	50.0	6	16.7
2	4	11.1	7	19.4
3	3	8.3	18	50.0
4	1	2.8	0	0.0
Total	36	100.0	36	100.0

Mean = 1.08
SD = 1.00

Mean = 2.06
SD = 1.12

Table 46.

Descriptive Statistics for the Main Effect of Number of Participants on the Number of Debriefed Posttest Errors for Unit #2

<u>Number of Debriefed Posttest Errors</u>	<u>Frequency (One-to-One)</u>	<u>Percent</u>	<u>Frequency (Small Group)</u>	<u>Percent</u>
0	1	2.8	0	0.0
1	7	19.4	0	0.0
2	4	11.1	5	13.9
3	11	30.6	6	16.7
4	2	5.6	18	50.0
5	7	19.4	6	16.7
6	3	8.3	0	0.0
7	1	2.8	0	0.0
missing	0	0.0	1	2.8
Total	36	100.0	36	100.0

Mean = 3.22
SD = 1.80

Mean = 4.57
SD = 1.22

Table 47.

Descriptive Statistics for the Main Effect of Number of Participants on the Number of Debriefed Posttest Errors in Total

<u>Number of Debriefed Posttest Errors</u>	<u>Frequency (One-to-One)</u>	<u>Percent</u>	<u>Frequency (Small Group)</u>	<u>Percent</u>
0	1	2.8	0	0.0
2	8	22.2	6	16.7
3	4	11.1	0	0.0
4	9	25.0	0	0.0
5	3	8.3	0	0.0
6	5	13.9	12	33.3
7	4	11.1	0	0.0
8	1	2.8	12	33.3
9	0	0.0	6	16.7
10	1	2.8	0	0.0
Total	36	100.0	36	100.0

Mean = 4.31
SD = 2.14

Mean = 6.50
SD = 2.32

Table 48.

Descriptive Statistics for the Interactions of Role and Number of Participants on the Number of Debriefed Posttest Errors for Unit #1

<u>Number of Posttest Errors Debriefed</u>	<u>Freq (#1)</u>	<u>%</u>	<u>Freq (#2)</u>	<u>%</u>	<u>Freq (#3)</u>	<u>%</u>	<u>Freq (#4)</u>	<u>%</u>
0	5	27.8	5	27.8	5	27.8	0	0.0
1	7	38.9	11	61.1	6	33.3	0	0.0
2	3	16.7	1	5.6	1	5.6	6	33.3
3	2	11.1	1	5.6	6	33.3	12	66.7
4	1	5.6	0	0.0	0	0.0	0	0.0
Total	18	100.0	18	100.0	18	100.0	18	100.0
Mean = 1.28 Mean = 0.89 Mean = 1.44 Mean = 2.67								
SD = 1.18 SD = 0.76 SD = 1.25 SD = 0.49								

Table 49.

Descriptive Statistics for the Interactions of Role and Number of Participants on the Number of Debriefed Posttest Errors for Unit #2

<u>Number of Posttest Errors Debriefed</u>	<u>Freq (#1)</u>	<u>%</u>	<u>Freq (#2)</u>	<u>%</u>	<u>Freq (#3)</u>	<u>%</u>	<u>Freq (#4)</u>	<u>%</u>
0	1	5.6	0	0.0	0	0.0	0	0.0
1	1	5.6	6	33.3	0	0.0	0	0.0
2	2	11.1	2	11.1	5	27.8	0	0.0
3	6	33.3	5	27.8	0	0.0	0	0.0
4	2	11.1	0	0.0	0	0.0	6	33.3
5	4	22.2	3	16.7	12	66.7	6	33.3
6	2	11.1	1	5.6	0	0.0	6	33.3
7	0	0.0	1	5.6	0	0.0	0	0.0
missing	0	0.0	0	0.0	1	5.6	0	0.0
Total	18	100.0	18	100.0	18	100.0	18	100.0
Mean = 3.50 Mean = 2.94 Mean = 4.12 Mean = 5.00								
SD = 1.65 SD = 1.92 SD = 1.41 SD = 0.84								

- * Key: *1 = One-to-One Active
 *2 = One-to-One Passive
 *3 = Small Group Active
 *4 = Small Group Passive

Table 50.

Descriptive Statistics for the Interactions of Role and Number of Participants on the Number of Debriefed Posttest Errors in Total

<u>Number of Posttest Errors Debriefed</u>	<u>Freq (#1)</u>	<u>%</u>	<u>Freq (#2)</u>	<u>%</u>	<u>Freq (#3)</u>	<u>%</u>	<u>Freq (#4)</u>	<u>%</u>
0	1	5.6	0	0.0	0	0.0	0	0.0
2	1	5.6	7	38.9	6	33.3	0	0.0
3	2	11.1	2	11.1	0	0.0	0	0.0
4	5	27.8	4	22.2	0	0.0	0	0.0
5	2	11.1	1	5.6	0	0.0	0	0.0
6	2	11.1	3	16.7	6	33.3	6	33.3
7	4	22.2	0	0.0	0	0.0	0	0.0
8	1	5.6	0	0.0	6	33.3	6	33.3
9	0	0.0	0	0.0	0	0.0	6	33.3
10	0	0.0	1	5.6	0	0.0	0	0.0
Total	18	100.0	18	100.0	18	100.0	18	100.0

Mean = 4.78	Mean = 3.83	Mean = 5.33	Mean = 7.67
SD = 2.07	SD = 2.15	SD = 2.57	SD = 1.28

* **Key:** *1= One-to-One Active
 *2= One-to-One Passive
 *3= Small Group Active
 *4= Small Group Passive

Table 51.

Means Based on the Two-Way ANOVA of the Number of Debriefed Posttest Errors

Role of Participants	Number of Participants	Debriefed Posttest Errors		
		Unit #1	Unit #2	Total
—	One-to-One	1.08	3.22	4.31
—	Small Group	2.06	4.57	6.50
Active	—	1.36	3.80	5.06
Passive	—	1.78	3.97	5.75
Active	One-to-One	1.28	3.50	4.78
	Small Group	1.44	4.12	5.33
Passive	One-to-One	0.89	2.94	3.83
	Small Group	2.67	5.00	7.67

Table 53.

Descriptive Statistics for the Number of Debriefed Survey Items Unit #1 by the Total Sample of Test Subjects

<u>Number of Debriefed Survey Items</u>	<u>Frequency</u>	<u>Percent</u>
1	4	5.6
2	10	13.9
3	13	18.1
4	4	5.6
5	16	22.2
6	11	15.3
7	1	1.4
9	1	1.4
10	2	2.8
14	1	1.4
15	6	8.3
17	2	2.8
19	1	1.4
Total	72	100.0

Mean = 5.81

SD = 4.46

Table 54.

Descriptive Statistics for the Number of Debriefed Survey Items Unit #2 by the Total Sample of Test Subjects

<u>Number of Debriefed Survey Items</u>	<u>Frequency</u>	<u>Percent</u>
0	1	1.4
1	5	6.9
2	18	25.0
3	3	4.2
4	2	2.8
5	10	13.9
6	8	11.1
7	7	9.7
9	2	2.8
10	7	9.7
14	6	8.3
15	1	1.4
17	1	1.4
19	1	1.4
Total	72	100.0

Mean = 6.00

SD = 4.78

Table 55.

Descriptive Statistics for the Number of Debriefed Survey Items in Total by the Total Sample of Test Subjects

<u>Number of Debriefed Survey Items</u>	<u>Frequency</u>	<u>Percent</u>
2	1	1.4
3	3	4.2
4	10	13.9
5	5	6.9
6	3	4.2
7	2	2.8
8	3	4.2
9	3	4.2
10	7	9.7
11	6	8.3
14	3	4.2
15	7	9.7
17	8	11.1
20	3	4.2
23	6	8.3
34	1	1.4
36	1	1.4
Total	72	100.0

Mean = 11.75

SD = 7.26

Table 56.

Descriptive Statistics for the Main Effect of Role of Participants on the Number of Debriefed Survey Items for Unit # 1

<u>Number of Debriefed Survey Items</u>	<u>Frequency (Active Role)</u>	<u>Percent</u>	<u>Frequency (Passive Role)</u>	<u>Percent</u>
1	2	5.6	2	5.6
2	8	22.2	2	5.6
3	3	8.3	10	27.8
4	2	5.6	2	5.6
5	10	27.8	6	16.7
6	1	2.8	10	27.8
7	1	2.8	0	0.0
9	1	2.8	0	0.0
10	0	0.0	2	5.6
14	1	2.8	0	0.0
15	6	16.7	0	0.0
17	1	2.8	1	2.8
19	0	0.0	1	2.8
Total	36	100.0	36	100.0

Mean = 6.33
SD = 5.07

Mean = 5.28
SD = 3.74

Table 57.

Descriptive Statistics for the Main Effect of Role of Participants on the Number of Debriefed Survey Items for Unit #2

<u>Number of Debriefed Survey Items</u>	<u>Frequency (Active Role)</u>	<u>Percent</u>	<u>Frequency (Passive Role)</u>	<u>Percent</u>
0	0	0.0	1	2.8
1	3	8.3	2	5.6
2	9	25.0	9	25.0
3	3	8.3	0	0.0
4	1	2.8	1	2.8
5	9	25.0	1	2.8
6	2	5.6	6	16.7
8	6	16.7	1	2.8
9	0	0.0	0	0.0
10	1	2.8	1	2.8
11	1	2.8	6	16.7
12	0	0.0	6	16.7
15	0	0.0	1	2.8
30	0	0.0	1	2.8
missing	1	2.8	0	0.0
Total	36	100.0	36	100.0

Mean = 4.57
SD = 2.73

Mean = 7.39
SD = 5.88

Table 58.

Descriptive Statistics for the Main Effect of Role of Participants on the Number of Debriefed Survey Items in Total

<u>Number of Debriefed Survey Items</u>	<u>Frequency (Active Role)</u>	<u>Percent</u>	<u>Frequency (Passive Role)</u>	<u>Percent</u>
2	0	0.0	1	2.8
3	2	5.6	1	2.8
4	7	19.4	3	8.3
5	2	5.6	3	8.3
6	2	5.6	1	2.8
7	2	5.6	0	0.0
8	1	2.8	2	5.6
9	2	5.6	1	2.8
10	7	19.4	0	0.0
11	0	0.0	6	16.7
14	1	2.8	2	5.6
15	1	2.8	6	16.7
17	1	2.8	7	19.4
20	2	5.6	1	2.8
23	6	16.7	0	0.0
34	0	0.0	1	2.8
36	0	0.0	1	2.8
Total	36	100.0	36	100.0

Mean = 10.83
SD = 7.04

Mean = 12.67
SD = 7.46

Table 59.

Descriptive Statistics for the Main Effect of Number of Participants on the Number of Debriefed Survey Items for Unit #1

<u>Number of Debriefed Survey Items</u>	<u>Frequency (One-to-One)</u>	<u>Percent</u>	<u>Frequency (Small Group)</u>	<u>Percent</u>
1	4	11.1	0	0.0
2	4	11.1	6	16.7
3	7	19.4	6	16.7
4	4	11.1	0	0.0
5	4	11.1	12	33.3
6	5	13.9	6	16.7
7	1	2.8	0	0.0
9	1	2.8	0	0.0
10	2	5.6	0	0.0
14	1	2.8	0	0.0
15	0	0.0	6	16.7
17	2	5.6	0	0.0
19	1	2.8	0	0.0
Total	36	100.0	36	100.0

Mean = 5.61
SD = 4.66

Mean = 6.00
SD = 4.30

Table 60.

Descriptive Statistics for the Main Effect of Number of Participants on the Number of Debriefed Survey Items for Unit #2

<u>Number of Debriefed Survey Items</u>	<u>Frequency (One-to-One)</u>	<u>Percent</u>	<u>Frequency (Small Group)</u>	<u>Percent</u>
0	1	2.8	0	0.0
1	5	13.9	0	0.0
2	13	36.1	5	13.9
3	3	8.3	0	0.0
4	2	5.6	0	0.0
5	4	11.1	6	16.7
6	2	5.6	6	16.7
8	1	2.8	6	16.7
9	0	0.0	0	0.0
10	2	5.6	0	0.0
11	1	2.8	6	16.7
12	0	0.0	6	16.7
15	1	2.8	0	0.0
30	1	2.8	0	0.0
missing	0	0.0	1	2.8
Total	36	100.0	36	100.0

Mean = 4.56
SD = 5.48

Mean = 7.49
SD = 3.43

Table 61.

Descriptive Statistics for the Main Effect of Number of Participants on the Number of Debriefed Survey
Items in Total

<u>Number of Debriefed Survey Items</u>	<u>Frequency (One-to-One)</u>	<u>Percent</u>	<u>Frequency (Small Group)</u>	<u>Percent</u>
2	1	2.8	0	0.0
3	3	8.3	0	0.0
4	4	11.1	6	16.7
5	5	13.9	0	0.0
6	3	8.3	0	0.0
7	2	5.6	0	0.0
8	3	8.3	0	0.0
9	3	8.3	0	0.0
10	1	2.8	6	16.7
11	0	0.0	6	16.7
14	3	8.3	0	0.0
15	1	2.8	6	16.7
17	2	5.6	6	16.7
20	3	8.3	0	0.0
23	0	0.0	6	16.7
34	1	2.8	0	0.0
36	1	2.8	0	0.0
Total	36	100.0	36	100.0

Mean = 10.17
SD = 8.08

Mean = 13.33
SD = 6.05

Table 62.

Descriptive Statistics for the Interactions of Role and Number of Participants on the Number of Debriefed Survey Items for Unit #1

<u>Number of Debriefed Survey Items</u>	<u>Freq (#1)</u>	<u>%</u>	<u>Freq (#2)</u>	<u>%</u>	<u>Freq (#3)</u>	<u>%</u>	<u>Freq (#4)</u>	<u>%</u>
1	2	11.1	2	11.1	0	0.0	0	0.0
2	2	11.1	2	11.1	6	33.3	0	0.0
3	3	16.7	4	22.2	0	0.0	6	33.3
4	2	11.1	2	11.1	0	0.0	0	0.0
5	4	22.2	0	0.0	6	33.3	6	33.3
6	1	5.6	4	22.2	0	0.0	6	33.3
7	1	5.6	0	0.0	0	0.0	0	0.0
9	1	5.6	0	0.0	0	0.0	0	0.0
10	0	0.0	2	11.1	0	0.0	0	0.0
14	1	5.6	0	0.0	0	0.0	0	0.0
15	0	0.0	0	0.0	6	33.3	0	0.0
17	1	5.6	1	5.6	0	0.0	0	0.0
19	0	0.0	1	5.6	0	0.0	0	0.0
Total	18	100.0	18	100.0	18	100.0	18	100.0
	Mean = 5.33		Mean = 5.89		Mean = 7.33		Mean = 4.67	
	SD = 4.26		SD = 5.13		SD = 5.72		SD = 1.28	

* Key: #1 = One-to-One Active
 #2 = One-to-One Passive
 #3 = Small Group Active
 #4 = Small Group Passive

Table 63.

Descriptive Statistics for the Interactions of Role and Number of Participants on the Number of Debriefed Survey Items for Unit #2

<u>Number of Debriefed Survey Items</u>	<u>Freq (#1)</u>	<u>%</u>	<u>Freq (#2)</u>	<u>%</u>	<u>Freq (#3)</u>	<u>%</u>	<u>Freq (#4)</u>	<u>%</u>
0	0	0.0	1	5.6	0	0.0	0	0.0
1	3	16.7	2	11.1	0	0.0	0	0.0
2	4	22.2	9	50.0	5	27.8	0	0.0
3	3	16.7	1	5.6	0	0.0	0	0.0
4	1	5.6	0	0.0	0	0.0	0	0.0
5	3	16.7	1	5.6	6	33.3	0	0.0
6	2	11.1	1	5.6	0	0.0	6	33.3
7	0	0.0	1	5.6	0	0.0	0	0.0
8	0	0.0	0	0.0	6	33.3	0	0.0
10	1	5.6	0	0.0	0	0.0	0	0.0
11	1	5.6	1	5.6	0	0.0	6	33.3
12	0	0.0	1	5.6	0	0.0	6	33.3
15	0	0.0	0	0.0	0	0.0	0	0.0
30	0	0.0	0	0.0	0	0.0	0	0.0
missing	0	0.0	0	0.0	1	5.6	0	0.0
Total	18	100.0	18	100.0	18	100.0	18	100.0
	Mean = 4.00		Mean = 5.11		Mean = 5.18		Mean = 9.67	
	SD = 2.89		SD = 7.27		SD = 2.48		SD = 2.70	

* Key: #1 = One-to-One Active
 #2 = One-to-One Passive
 #3 = Small Group Active
 #4 = Small Group Passive

Table 64.

Descriptive Statistics for the Interactions of Role and Number of Participants on the Number of Debriefed Survey Items in Total

<u>Number of Debriefed Survey Items</u>	<u>Freq (#1)</u>	<u>%</u>	<u>Freq (#2)</u>	<u>%</u>	<u>Freq (#3)</u>	<u>%</u>	<u>Freq (#4)</u>	<u>%</u>
2	0	0.0	1	5.6	0	0.0	0	0.0
3	2	11.1	1	5.6	0	0.0	0	0.0
4	1	5.6	3	16.7	6	33.3	0	0.0
5	2	11.1	3	16.7	0	0.0	0	0.0
6	2	11.1	1	5.6	0	0.0	0	0.0
7	2	11.1	0	0.0	0	0.0	0	0.0
8	1	5.6	2	11.1	0	0.0	0	0.0
9	2	11.1	1	5.6	0	0.0	0	0.0
10	0	0.0	0	0.0	6	33.3	0	0.0
11	1	5.6	0	0.0	0	0.0	6	33.3
12	0	0.0	0	0.0	0	0.0	0	0.0
13	0	0.0	0	0.0	0	0.0	0	0.0
14	1	5.6	2	11.1	0	0.0	0	0.0
15	1	5.6	0	0.0	0	0.0	6	33.3
17	1	5.6	1	5.6	0	0.0	6	33.3
20	2	11.1	1	5.6	0	0.0	0	0.0
23	0	0.0	0	0.0	6	33.3	0	0.0
34	0	0.0	1	5.6	0	0.0	0	0.0
36	0	0.0	1	5.6	0	0.0	0	0.0
Total	18	100.0	18	100.0	18	100.0	18	100.0

Mean = 9.33	Mean = 11.0	Mean = 12.33	Mean = 14.33
SD = 5.54	SD = 10.11	SD = 8.16	SD = 2.57

* Key: #1 = One-to-One Active
 #2 = One-to-One Passive
 #3 = Small Group Active
 #4 = Small Group Passive

Table 65.

Means Based on the Two-Way ANOVA of the Number of Debriefed Survey Items

Role of Participants	Number of Participants	Debriefed Survey Items		
		Unit #1	Unit #2	Total
—	One-to-One	5.61	4.56	10.17
	Small Group	6.00	7.49	13.33
Active	—	6.33	4.57	10.83
Passive	—	5.28	7.39	12.67
Active Passive	One-to-One	5.33	4.00	9.33
	Small Group	7.33	5.18	12.33
	One-to-One	5.89	5.11	11.00
	Small Group	4.67	9.67	14.33

Table 67.

Descriptive Statistics for the Total Number of Agenda Items in Total for the Total Sample of Test Subjects

<u>Total Number of Agenda Items</u>	<u>Frequency</u>	<u>Percent</u>
4	1	1.4
5	2	2.8
6	7	9.7
8	2	2.8
9	1	1.4
10	2	2.8
11	2	2.8
12	3	4.2
13	3	4.2
14	3	4.2
15	1	1.4
18	1	1.4
20	6	8.3
21	1	1.4
22	7	9.7
24	1	1.4
25	1	1.4
27	6	8.3
28	1	1.4
29	7	9.7
30	1	1.4
31	7	9.7
32	1	1.4
40	1	1.4
41	1	1.4
42	1	1.4
50	1	1.4
64	1	1.4
Total	72	100.0

Mean = 21.19

SD = 11.55

Table 68.

Descriptive Statistics for the Main Effect of Role of Participants on the Total Number of Agenda Items in
Total

<u>Total Number of Agenda</u> <u>Items</u>	<u>Frequency</u> <u>(Active Role)</u>	<u>Percent</u>	<u>Frequency</u> <u>(Passive Role)</u>	<u>Percent</u>
4	0	0.0	1	2.8
5	1	2.8	1	2.8
6	6	16.7	1	2.8
8	1	2.8	1	2.8
9	0	0.0	1	2.8
10	1	2.8	1	2.8
11	0	0.0	2	5.6
12	2	5.6	1	2.8
13	2	5.6	1	2.8
14	2	5.6	1	2.8
15	1	2.8	0	0.0
18	1	2.8	0	0.0
20	6	16.7	0	0.0
21	1	2.8	0	0.0
22	0	0.0	7	19.4
24	1	2.8	0	0.0
25	1	2.8	0	0.0
27	0	0.0	6	16.7
28	1	2.8	0	0.0
29	1	2.8	6	16.7
30	0	0.0	1	2.8
31	7	19.4	0	0.0
32	0	0.0	1	2.8
40	0	0.0	1	2.8
41	0	0.0	1	2.8
42	0	0.0	1	2.8
50	0	0.0	1	2.8
64	1	2.8	0	0.0
Total	36	100.0	36	100.0

Mean = 19.39
SD = 11.86

Mean = 23.00
SD = 11.09

Table 69.

Descriptive Statistics for the Main Effect of Number of Participants on the Total Number of Agenda
Items in Total

<u>Total Number of Agenda Items</u>	<u>Frequency (One-to-One)</u>	<u>Percent</u>	<u>Frequency (Small Group)</u>	<u>Percent</u>
4	1	2.8	0	0.0
5	2	5.6	0	0.0
6	1	2.8	6	16.7
8	2	5.6	0	0.0
9	1	2.8	0	0.0
10	2	5.6	0	0.0
11	2	5.6	0	0.0
12	3	8.3	0	0.0
13	3	8.3	0	0.0
14	3	8.3	0	0.0
15	1	2.8	0	0.0
18	1	2.8	0	0.0
20	0	0.0	6	16.7
21	1	2.8	0	0.0
22	1	2.8	6	16.7
24	1	2.8	0	0.0
25	1	2.8	0	0.0
27	0	0.0	6	16.7
28	1	2.8	0	0.0
29	1	2.8	6	16.7
30	1	2.8	0	0.0
31	1	2.8	6	16.7
32	1	2.8	0	0.0
40	1	2.8	0	0.0
41	1	2.8	0	0.0
42	1	2.8	0	0.0
50	1	2.8	0	0.0
64	1	2.8	0	0.0
Total	36	100.0	36	100.0

Mean = 19.89
SD = 14.01

Mean = 22.50
SD = 8.42

Table 70.

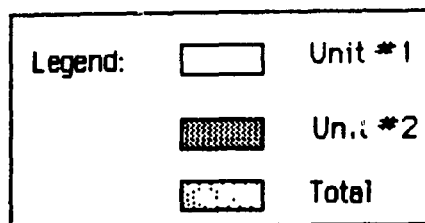
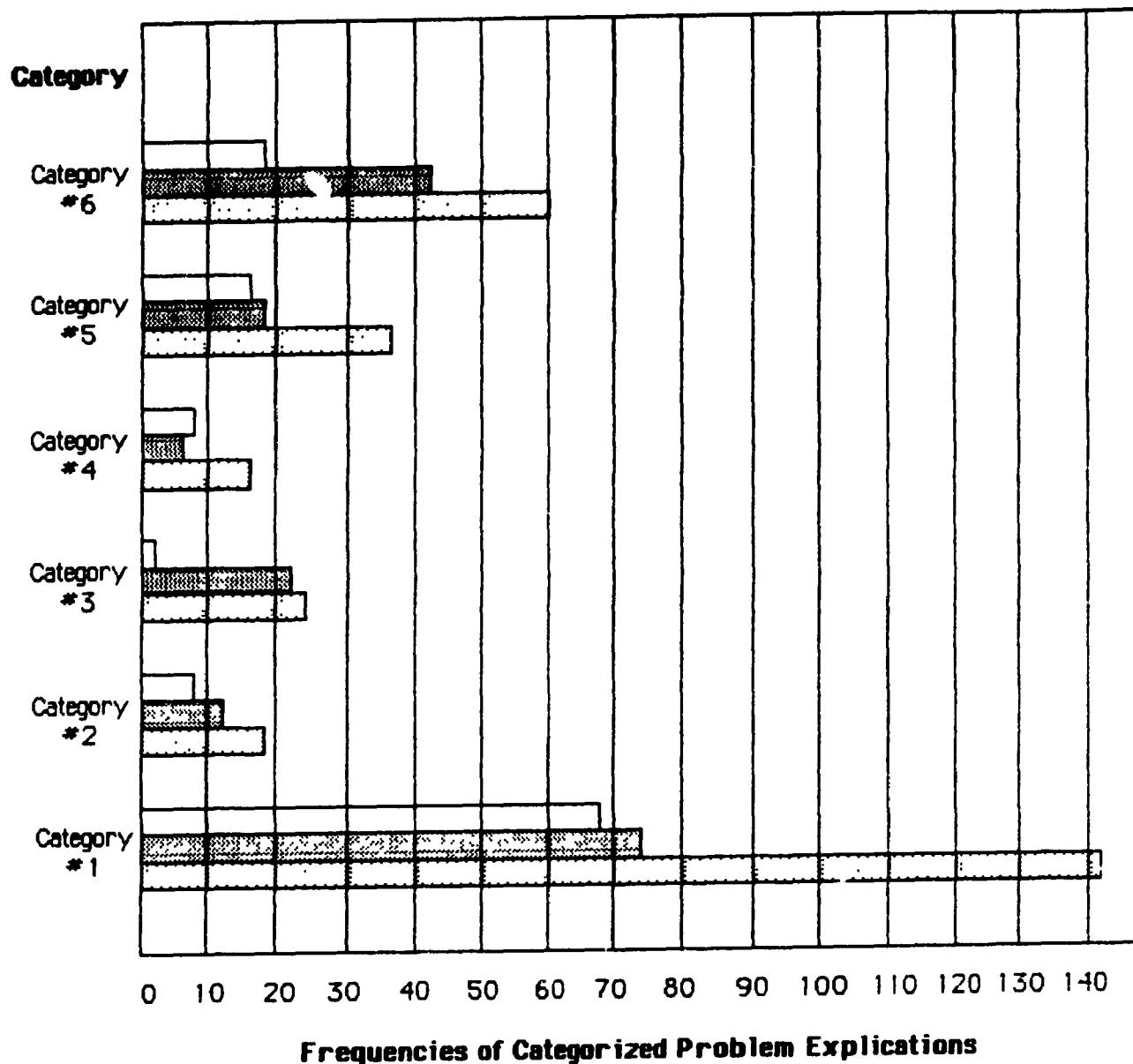
Descriptive Statistics for the Interactions of Role and Number of Participants on the Total Number of Agenda Items in Total

<u>Total Number of Agenda Items</u>	<u>Freq (#1)</u>	<u>%</u>	<u>Freq (#2)</u>	<u>%</u>	<u>Freq (#3)</u>	<u>%</u>	<u>Freq (#4)</u>	<u>%</u>
4	0	0.0	1	5.6	0	0.0	0	0.0
5	1	5.6	1	5.6	0	0.0	0	0.0
6	0	0.0	1	5.6	6	33.3	0	0.0
8	1	5.6	1	5.6	0	0.0	0	0.0
9	0	0.0	1	5.6	0	0.0	0	0.0
10	1	5.6	1	5.6	0	0.0	0	0.0
11	0	0.0	2	11.1	0	0.0	0	0.0
12	2	11.1	1	5.6	0	0.0	0	0.0
13	2	11.1	1	5.6	0	0.0	0	0.0
14	2	11.1	1	5.6	0	0.0	0	0.0
15	1	5.6	0	0.0	0	0.0	0	0.0
18	1	5.6	0	0.0	0	0.0	0	0.0
20	0	0.0	0	0.0	6	33.3	0	0.0
21	1	5.6	0	0.0	0	0.0	0	0.0
22	0	0.0	1	5.6	0	0.0	6	33.3
24	1	5.6	0	0.0	0	0.0	0	0.0
25	1	5.6	0	0.0	0	0.0	0	0.0
27	0	0.0	0	0.0	0	0.0	6	33.3
28	1	5.6	0	0.0	0	0.0	0	0.0
29	1	5.6	0	0.0	0	0.0	6	33.3
30	0	0.0	1	5.6	0	0.0	0	0.0
31	1	5.6	0	0.0	6	33.3	0	0.0
32	0	0.0	1	5.6	0	0.0	0	0.0
39	0	0.0	0	0.0	0	0.0	0	0.0
40	0	0.0	1	5.6	0	0.0	0	0.0
41	0	0.0	1	5.6	0	0.0	0	0.0
42	0	0.0	1	5.6	0	0.0	0	0.0
50	0	0.0	1	5.6	0	0.0	0	0.0
64	1	5.6	0	0.0	0	0.0	0	0.0
Total	18	100.0	18	100.0	18	100.0	18	100.0

Mean = 19.78 Mean = 10.00 Mean = 19.00 Mean = 26.00
SD = 13.37 SD = 15.00 SD = 10.53 SD = 3.03

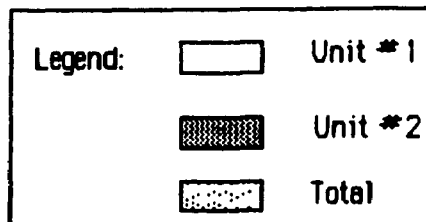
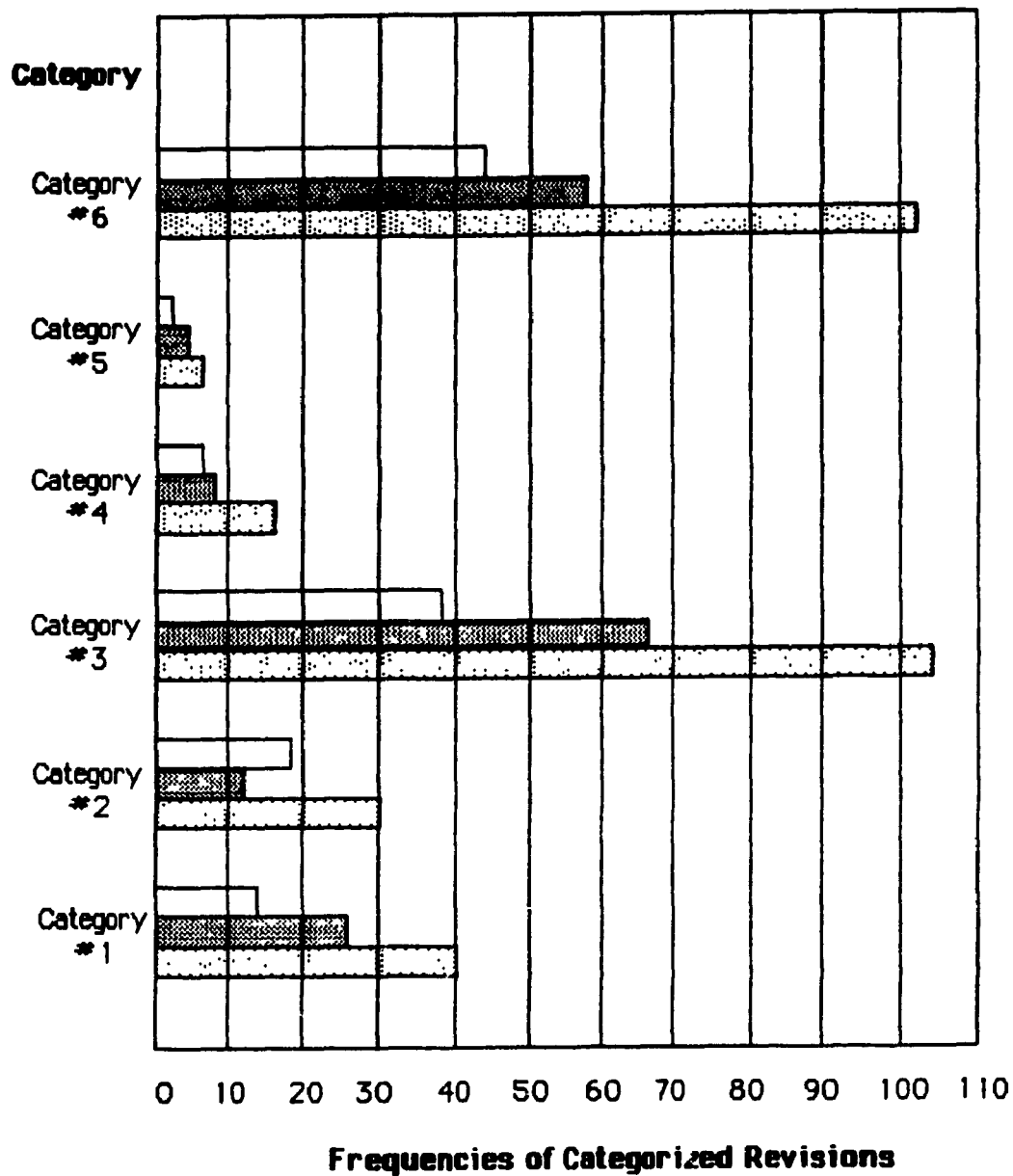
* Key: #1 = One-to-One Active
#2 = One-to-One Passive
#3 = Small Group Active
#4 = Small Group Passive

APPENDIX J



Category #1 = Comprehension/Recall
 Category #2 = Motivation/Attention
 Category #3 = Physical Features
 Category #4 = Multiple Responses Overlapping
 Two or More of the Above
 Category #5 = Non-Classifiable Responses
 Category #6 = Problem Located but not
 Explicated

Figure 19. A histogram showing frequencies of categorized problem explanations by unit for the total sample (N=72). Note that unit #2 frequencies are based on (N=71).



Category #1 = Comprehension/Recall
 Category #2 = Motivation/Attention
 Category #3 = Physical Features
 Category #4 = Multiple Responses Overlapping
 Two or More of the Above
 Category #5 = Non-Classifiable Responses
 Category #6 = Problem Located but not
 Revised

Figure 20. A histogram showing frequencies of categorized revisions by unit for the total sample (N=72). Note that unit #2 frequencies are based on (N=71).

Table 74.

Weighted Means Based on the 2 X 2 MANOVA of Categorized Problem Explications for Unit #1

Role of Participants	Number of Participants	Categories of Problem Explications					
		#1	#2	#3	#4	#5	#6
Active	One-to-One	1.33	0.22	0.00	0.33	0.39	0.00
	Small Group	0.17	0.00	0.00	0.00	0.11	0.00
Passive	One-to-One	1.22	0.17	0.00	0.11	0.28	0.78
	Small Group	1.00	0.06	0.06	0.00	0.17	0.22
Active	—	0.75	0.11	0.00	0.17	0.25	0.00
Passive	—	1.11	0.11	0.03	0.06	0.22	0.50
—	One-to-One	1.28	0.19	0.00	0.22	0.33	0.39
—	Small Group	0.58	0.03	0.03	0.00	0.14	0.11

Table 75.

Weighted Means Based on the 2 X 2 MANOVA of Categorized Revisions for Unit #1

Role of Participants	Number of Participants	Categories of Revisions					
		#1	#2	#3	#4	#5	#6
Active	One-to-One	0.22	0.61	0.67	0.17	0.06	0.56
	Small Group	0.11	0.00	0.17	0.00	0.00	0.00
Passive	One-to-One	0.11	0.06	0.89	0.11	0.00	1.39
	Small Group	0.33	0.33	0.39	0.00	0.00	0.44
Active	—	0.17	0.31	0.42	0.08	0.03	0.28
Passive	—	0.22	0.19	0.64	0.06	0.00	0.92
—	One-to-One	0.17	0.33	0.78	0.14	0.03	0.97
—	Small Group	0.22	0.17	0.28	0.00	0.00	0.22

Table 78.

Weighted Means Based on the 2 X 2 MANOVA of Categorized Problem Explications for Unit #2

Role of Participants	Number of Participants	Categories of Problem Explications					
		#1	#2	#3	#4	#5	#6
Active	One-to-One	1.94	0.22	0.33	0.22	0.50	0.00
	Small Group	0.24	0.06	0.18	0.06	0.12	0.12
Passive	One-to-One	0.61	0.00	0.67	0.00	0.39	0.83
	Small Group	1.33	0.33	0.11	0.11	0.00	1.39
Active	—	1.11	0.14	0.26	0.14	0.31	0.06
Passive	—	0.97	0.17	0.39	0.06	0.19	1.11
—	One-to-One	1.28	0.11	0.50	0.11	0.44	0.42
—	Small Group	0.80	0.20	0.14	0.09	0.06	0.77

Table 79.

Weighted Means Based on the 2 X 2 MANOVA of Categorized Revisions for Unit #2

Role of Participants	Number of Participants	Categories of Revisions					
		#1	#2	#3	#4	#5	#6
Active	One-to-One	0.83	0.22	1.50	0.22	0.17	0.28
	Small Group	0.06	0.00	0.35	0.06	0.00	0.29
Passive	One-to-One	0.33	0.00	0.56	0.00	0.00	1.61
	Small Group	0.22	0.44	1.22	0.22	0.06	1.11
Active	—	0.46	0.11	0.94	0.14	0.09	0.29
Passive	—	0.28	0.22	0.89	0.11	0.03	1.36
—	One-to-One	0.58	0.11	1.03	0.11	0.08	0.94
—	Small Group	0.14	0.23	0.80	0.14	0.03	0.71

Table 83.

Weighted Means Based on the 2 X 2 MANOVA of the Total Number of Categorized Problem Explanations

Role of Participants	Number of Participants	Categories of Problem Explanations					
		#1	#2	#3	#4	#5	#6
Active	One-to-One	3.28	0.44	0.33	0.56	0.89	0.00
	Small Group	0.39	0.06	0.17	0.06	0.22	0.11
Passive	One-to-One	1.83	0.17	0.67	0.11	0.67	1.61
	Small Group	2.33	0.39	0.17	0.11	0.17	1.61
Active	—	1.83	0.25	0.25	0.31	0.56	0.06
Passive	—	2.08	0.28	0.42	0.11	0.42	1.61
—	One-to-One	2.56	0.31	0.50	0.33	0.78	0.81
—	Small Group	1.36	0.22	0.17	0.08	0.19	0.86

Table 84.

Weighted Means Based on the 2 X 2 MANOVA of the Total Number of Categorized Revisions

Role of Participants	Number of Participants	Categories of Revisions					
		#1	#2	#3	#4	#5	#6
Active	One-to-One	1.06	0.83	2.17	0.39	0.22	0.83
	Small Group	0.17	0.00	0.50	0.06	0.00	0.28
Passive	One-to-One	0.44	0.06	1.44	0.11	0.00	3.00
	Small Group	0.56	0.78	1.61	0.22	0.06	1.56
Active	—	0.61	0.42	1.33	0.22	0.11	0.56
Passive	—	0.50	0.42	1.53	0.17	0.03	2.28
—	One-to-One	0.75	0.44	1.81	0.25	0.11	1.92
—	Small Group	0.36	0.39	1.06	0.14	0.03	0.91

APPENDIX K

Table 87.

Weighted Means Based on the 2 X 2 X 2 MANOVA of Categorized Problem Explanations for Unit #1

Role of Participants	Prior Knowledge	Number of Participants	Categories of Problem Explanations:					
			#1	#2	#3	#4	#5	#6
Active	—	—	0.77	0.11	0.00	0.17	0.26	0.00
Passive	—	—	1.11	0.11	0.03	0.06	0.22	0.50
—	Prior	—	0.56	0.03	0.00	0.03	0.29	0.26
—	No Prior	—	1.30	0.19	0.03	0.19	0.19	0.24
—	—	One-to-One	1.28	0.19	0.00	0.22	0.33	0.39
—	—	Small Group	0.60	0.03	0.03	0.00	0.14	0.11
Active	Prior	—	0.25	0.00	0.00	0.00	0.30	0.00
	No Prior	—	1.47	0.27	0.00	0.40	0.20	0.00
Passive	Prior	—	1.00	0.07	0.00	0.07	0.29	0.64
	No Prior	—	1.18	0.14	0.05	0.05	0.18	0.41
Active	—	One-to-One	1.33	0.22	0.00	0.33	0.39	0.00
	—	Small Group	0.18	0.00	0.00	0.00	0.12	0.00
Passive	—	One-to-One	1.22	0.17	0.00	0.11	0.28	0.78
	—	Small Group	1.00	0.06	0.06	0.00	0.17	0.22
—	Prior	One-to-One	0.72	0.06	0.00	0.06	0.44	0.50
		Small Group	0.38	0.00	0.00	0.00	0.13	0.00
	No Prior	One-to-One	1.83	0.33	0.00	0.39	0.22	0.28
		Small Group	0.79	0.05	0.05	0.00	0.16	0.21
Active	Prior	One-to-One	0.44	0.00	0.00	0.00	0.56	0.00
		Small Group	0.09	0.00	0.00	0.00	0.09	0.00
	No Prior	One-to-One	2.22	0.44	0.00	0.67	0.22	0.00
		Small Group	0.33	0.00	0.00	0.00	0.17	0.00
Passive	Prior	One-to-One	1.00	0.11	0.00	0.11	0.33	1.00
		Small Group	1.00	0.00	0.00	0.00	0.20	0.00
	No Prior	One-to-One	1.44	0.22	0.00	0.11	0.22	0.56
		Small Group	1.00	0.08	0.08	0.00	0.15	0.31

Table 88.

Weighted Means Based on the 2 X 2 X 2 MANOVA of Categorized Revisions for Unit #1

Role of Participants	Prior Knowledge	Number of Participants	Categories of Revisions:					
			#1	#2	#3	#4	#5	#6
Active	—	—	0.17	0.31	0.43	0.09	0.03	0.29
Passive	—	—	0.22	0.19	0.64	0.06	0.00	0.92
—	Prior	—	0.15	0.00	0.32	0.03	0.00	0.68
—	No Prior	—	0.24	0.49	0.73	0.11	0.03	0.54
—	—	One-to-One	0.17	0.33	0.78	0.14	0.03	0.97
—	—	Small Group	0.23	0.17	0.29	0.00	0.00	0.23
Active	Prior	—	0.10	0.00	0.15	0.00	0.00	0.30
	No Prior	—	0.27	0.73	0.80	0.20	0.07	0.27
Passive	Prior	—	0.21	0.00	0.57	0.07	0.00	1.21
	No Prior	—	0.23	0.32	0.68	0.05	0.00	0.73
Active	—	One-to-One	0.22	0.61	0.67	0.17	0.06	0.56
	—	Small Group	0.12	0.00	0.17	0.00	0.00	0.00
Passive	—	One-to-One	0.11	0.06	0.89	0.11	0.00	1.39
	—	Small Group	0.33	0.33	0.39	0.00	0.00	0.44
—	Prior	One-to-One	0.17	0.00	0.44	0.06	0.00	1.11
		Small Group	0.13	0.00	0.19	0.00	0.00	0.19
	No Prior	One-to-One	0.17	0.67	1.11	0.22	0.06	0.83
		Small Group	0.32	0.32	0.37	0.00	0.00	0.26
Active	Prior	One-to-One	0.11	0.00	0.22	0.00	0.00	0.67
		Small Group	0.09	0.00	0.09	0.00	0.00	0.00
	No Prior	One-to-One	0.33	1.22	1.11	0.33	0.11	0.44
		Small Group	0.17	0.00	0.33	0.00	0.00	0.00
	Prior	One-to-One	0.22	0.00	0.67	0.11	0.00	1.56
		Small Group	0.20	0.00	0.40	0.00	0.00	0.60
	No Prior	One-to-One	0.00	0.11	1.11	0.11	0.00	1.22
		Small Group	0.38	0.46	0.38	0.00	0.00	0.38

Table 91.

Weighted Means Based on the 2 X 2 X 2 MANOVA of Categorized Problem Explications
for Unit #2

Role of Participants	Prior Knowledge	Number of Participants	Categories of Problem Explications:					
			#1	#2	#3	#4	#5	#6
Active	—	—	1.11	0.14	0.26	0.14	0.31	0.06
Passive	—	—	0.97	0.17	0.39	0.06	0.19	1.11
—	Prior	—	0.53	0.09	0.26	0.00	0.12	0.44
—	No Prior	—	1.51	0.22	0.38	0.19	0.38	0.73
—	—	One-to-One	1.28	0.11	0.50	0.11	0.44	0.42
—	—	Small Group	0.80	0.20	0.14	0.09	0.06	0.77
Active	Prior	—	0.55	0.00	0.15	0.00	0.10	0.00
	No Prior	—	1.87	0.33	0.40	0.33	0.60	0.13
Passive	Prior	—	0.50	0.21	0.43	0.00	0.14	1.07
	No Prior	—	1.27	0.14	0.36	0.09	0.23	1.14
Active	—	One-to-One	1.94	0.22	0.33	0.22	0.50	0.00
	—	Small Group	0.24	0.06	0.18	0.06	0.12	0.12
Passive	—	One-to-One	0.61	0.00	0.67	0.00	0.39	0.83
	—	Small Group	1.33	0.33	0.11	0.11	0.00	1.39
—	Prior	One-to-One	0.89	0.00	0.50	0.00	0.22	0.72
		Small Group	0.13	0.19	0.00	0.00	0.00	0.13
	No Prior	One-to-One	1.67	0.22	0.50	0.22	0.67	0.11
		Small Group	1.37	0.21	0.26	0.16	0.11	1.32
Active	Prior	One-to-One	1.22	0.00	0.33	0.00	0.22	0.00
		Small Group	0.00	0.00	0.00	0.00	0.00	0.00
	No Prior	One-to-One	2.67	0.44	0.33	0.44	0.78	0.00
		Small Group	0.67	0.17	0.50	0.17	0.33	0.33
	Prior	One-to-One	0.56	0.00	0.67	0.00	0.22	1.44
		Small Group	0.40	0.60	0.00	0.00	0.00	0.40
	No Prior	One-to-One	0.67	0.00	0.67	0.00	0.56	0.22
		Small Group	1.70	0.23	0.15	0.15	0.00	1.77

Table 92.

Weighted Means Based on the 2 X 2 X 2 MANOVA of Categorized Revisions for Unit #2

Role of Participants	Prior Knowledge	Number of Participants	Categories of Revisions					
			#1	#2	#3	#4	#5	#6
Active	—	—	0.46	0.11	0.94	0.14	0.09	0.29
Passive	—	—	0.28	0.22	0.89	0.11	0.03	1.36
—	Prior	—	0.24	0.15	0.32	0.00	0.00	0.74
—	No Prior	—	0.49	0.19	1.46	0.24	0.11	0.92
—	—	One-to-One	0.58	0.11	1.03	0.11	0.08	0.94
—	—	Small Group	0.14	0.23	0.80	0.14	0.03	0.71
Active	Prior	—	0.35	0.10	0.25	0.00	0.00	0.10
	No Prior	—	0.60	0.13	1.87	0.33	0.20	0.53
Passive	Prior	—	0.07	0.21	0.43	0.00	0.00	1.64
	No Prior	—	0.41	0.23	1.18	0.18	0.05	1.18
Active	—	One-to-One	0.83	0.22	1.50	0.22	0.16	0.28
	—	Small Group	0.06	0.00	0.35	0.06	0.00	0.29
Passive	—	One-to-One	0.33	0.00	0.56	0.00	0.00	1.61
	—	Small Group	0.22	0.44	1.22	0.22	0.06	1.11
—	Prior	One-to-One	0.39	0.11	0.56	0.00	0.00	1.28
		Small Group	0.06	0.19	0.06	0.00	0.00	0.13
	No Prior	One-to-One	0.78	0.11	1.50	0.22	0.17	0.61
		Small Group	0.21	0.26	1.42	0.26	0.05	1.21
Active	Prior	One-to-One	0.78	0.22	0.56	0.00	0.00	0.22
		Small Group	0.00	0.00	0.00	0.00	0.00	0.00
	No Prior	One-to-One	0.89	0.22	2.44	0.44	0.33	0.33
		Small Group	0.17	0.00	1.00	0.17	0.00	0.83
	Prior	One-to-One	0.00	0.00	0.56	0.00	0.00	2.33
		Small Group	0.20	0.60	0.20	0.00	0.00	0.40
	No Prior	One-to-One	0.67	0.00	0.56	0.00	0.00	0.89
		Small Group	0.23	0.38	1.62	0.31	0.08	1.38

Table 95.

Weighted Means Based on the 2 X 2 X 2 MANOVA of the Total Number of Categorized Problem Explications

Role of Participants	Prior Knowledge	Number of Participants	Categories of Problem Explications:					
			#1	#2	#3	#4	#5	#6
Active	—	—	1.89	0.26	0.26	0.31	0.57	0.06
Passive	—	—	2.08	0.28	0.42	0.11	0.42	1.61
—	Prior	—	1.09	0.12	0.26	0.03	0.41	0.71
—	No Prior	—	2.81	0.41	0.41	0.38	0.57	0.97
—	—	One-to-One	2.56	0.31	0.50	0.33	0.78	0.81
—	—	Small Group	1.40	0.23	0.17	0.09	0.20	0.89
Active	Prior	—	0.80	0.00	0.15	0.00	0.40	0.00
	No Prior	—	3.33	0.60	0.40	0.73	0.80	0.13
Passive	Prior	—	1.50	0.29	0.43	0.07	0.43	1.71
	No Prior	—	2.45	0.27	0.41	0.14	0.41	1.55
Active	—	One-to-One	3.28	0.44	0.33	0.56	0.89	0.00
	—	Small Group	0.41	0.06	0.18	0.06	0.24	0.12
Passive	—	One-to-One	1.83	0.17	0.57	0.11	0.67	1.61
	—	Small Group	2.33	0.39	0.17	0.11	0.17	1.61
—	Prior	One-to-One	1.61	0.06	0.50	0.06	0.67	1.22
		Small Group	0.50	0.19	0.00	0.00	0.13	0.13
	No Prior	One-to-One	3.50	0.56	0.50	0.61	0.89	0.39
		Small Group	2.16	0.26	0.32	0.16	0.26	1.53
Active	Prior	One-to-One	1.67	0.00	0.33	0.00	0.78	0.00
		Small Group	0.09	0.00	0.00	0.00	0.09	0.00
	No Prior	One-to-One	4.89	0.89	0.33	1.11	1.00	0.00
		Small Group	1.00	0.17	0.50	0.17	0.50	0.33
	Prior	One-to-One	1.55	0.11	0.67	0.11	0.56	2.44
		Small Group	1.40	0.60	0.00	0.00	0.20	0.40
	No Prior	One-to-One	2.11	0.22	0.67	0.11	0.78	0.78
		Small Group	2.69	0.31	0.23	0.15	0.15	2.08

Table 96.

Weighted Means Based on the 2 X 2 X 2 MANOVA of the Total Number of Categorized Revisions

Role of Participants	Prior Knowledge	Number of Participants	Categories of Revisions:					
			#1	#2	#3	#4	#5	#6
Active	—	—	0.63	0.43	1.37	0.23	0.11	0.57
Passive	—	—	0.50	0.42	1.53	0.17	0.03	2.28
—	Prior	—	0.38	0.15	0.65	0.03	0.00	1.41
—	No Prior	—	0.73	0.68	2.19	0.35	0.14	1.46
—	—	One-to-One	0.75	0.44	1.81	0.25	0.11	1.92
—	—	Small Group	0.37	0.40	1.09	0.14	0.03	0.94
Active	Prior	—	0.45	0.10	0.40	0.00	0.00	0.40
	No Prior	—	0.87	0.87	2.67	0.53	0.27	0.80
Passive	Prior	—	0.29	0.21	1.00	0.07	0.00	2.86
	No Prior	—	0.64	0.55	1.86	0.23	0.05	1.91
Active	—	One-to-One	1.06	0.83	2.17	0.39	0.22	0.83
	—	Small Group	0.18	0.00	0.53	0.06	0.00	0.29
Passive	—	One-to-One	0.44	0.06	1.44	0.11	0.00	3.00
	--	Small Group	0.56	0.78	1.61	0.22	0.06	1.56
—	Prior	One-to-One	0.56	0.11	1.00	0.06	0.00	2.39
		Small Group	0.19	0.19	0.25	0.00	0.00	0.31
	No Prior	One-to-One	0.94	0.78	2.61	0.44	0.22	1.44
		Small Group	0.53	0.58	1.79	0.26	0.06	1.47
Active	Prior	One-to-One	0.89	0.22	0.78	0.00	0.00	0.89
		Small Group	0.09	0.00	0.09	0.00	0.00	0.00
	No Prior	One-to-One	1.22	1.44	3.56	0.78	0.44	0.78
		Small Group	0.33	0.00	1.33	0.17	0.00	0.83
	Prior	One-to-One	0.22	0.00	1.22	0.11	0.00	3.89
		Small Group	0.40	0.60	0.60	0.00	0.00	1.00
	No Prior	One-to-One	0.67	0.11	1.67	0.11	0.00	2.11
		Small Group	0.62	0.85	2.00	0.31	0.08	1.77