CHAPTER III

METHODOLOGY

Introduction

The previous chapter was used to establish the conceptual and operational relationships between evaluation, content analysis, and microcomputers. This chapter is used to describe an experimental study in which microcomputers are used to help pre-service and practicing educators perform a content analysis of responses to an open-ended survey question used in a simulated evaluation effort. The study is discussed under three main topics: (1) an overview of the study, (2) the procedures employed, and (3) the data analyses used.

Overview of the Study

This overview is used to link the concepts presented in Chapter 2 to the experimental study discussed here. It includes summaries of: (a) the problem, (b) the simulation activity on which the study is based, (c) the design of the study, (d) the independent variables, (e) the research hypotheses, and (f) the study group and sample.

Review of the Problem

Evaluation practitioners must often collect and analyze responses to a set of spoken or written questions obtained from large groups of people. These questions may be "forced-choice," in which

valid responses are determined in advance and the respondents must choose from this set of fixed responses; or, at the other extreme, the questions may be "open-ended," in which the questions are phrased to identify the topics of the desired responses but respondents are left to answer the questions in their own words.

Responses to forced-choice questions are usually "quantitative" and are best analyzed by using statistical analysis procedures of one sort or another. Responses to open-ended questions are usually "qualitative" and are best analyzed by using content analysis procedures. Both types of analysis require the use of specialized skills and "tools." Most evaluation practitioners are familiar with how to conduct fundamental statistical analysis procedures or at least have access to someone who is familiar with them. They also have access to the tools for statistical analysis--computer programs to obtain descriptive or inferential statistics and the equipment to run those programs.

On the other hand, practitioners conducting evaluations that require the analysis of many responses to open-ended questions often find themselves overwhelmed by the magnitude of the task. This problem occurs because they usually have little formal training in content analysis theories and methods and they have inadequate tools to perform the task. Inadequate skills can be upgraded by providing appropriate training about content analysis; but the computer program tools used in content analysis have traditionally been large, specialized, and expensive to operate, making them effectively unavailable to most evaluation practitioners.

Fortunately, certain content analysis techniques can now be adapted to work with general purpose programs running on relatively inexpensive microcomputers. In particular, word lists can be generated with some spelling checker programs, and information retrieval can be performed with data base management programs. This means techniques previously available only to content analysis experts with large and expensive mainframe computers running highly specialized programs are now available to evaluation practitioners with microcomputers and a set of general purpose programs.

Two fundamental content analysis tasks that can be implemented with the use of microcomputers include developing a category system and coding the set of responses in terms of that system. If using a microcomputer improves the quality of the analysis, or saves time and money while holding quality constant, it is a worthwhile investment. Two key concepts for judging the quality of a content analysis are reliability and validity. These standards of quality can be applied to both developing the categories and coding the responses in relation to them. They can also provide the basis for judging if using microcomputers produces high quality results.

The problem then becomes one of determining if available microcomputer programs can be used by evaluation practitioners to obtain high quality results when analyzing narrative survey responses. More formally, the general problem can be stated as follows: How can evaluation practitioners use microcomputer programs to obtain reliable and valid content analyses of responses to open-ended survey questions?

Simulation Activity

An ideal situation for addressing the above problem would be one in which a large number of practitioners independently analyzed a set of responses to an open-ended question used in a real-world evaluation. However, real evaluations and content analyses are never performed in this way. Instead, they are typically performed by only a few individuals. The solution to this difficulty is to create a simulation activity consistent with the general model for conducting an evaluation effort based on a non-trivial, actual evaluation that solicits responses to an open-ended question. Such an evaluation is described by Patton (1980, pp. 23-30), and the evaluation report (Patton, French, & Perrone, 1976) was used as the basis for developing the responses used in the simulation activity. This development process is summarized in the procedures section of this chapter. It is also discussed in more detail in Appendix A.

The simulation activity placed each study participant in the role of a student research assistant working at a university research center. Because the director was just called away to an important meeting, she asked the student to summarize a set of responses to an open-ended survey question used in an ongoing evaluation project. The project involved evaluating a controversial accountability system of a moderately large public school district from the perspective of the teachers. The student was asked to first develop a category system based on 50 teacher responses and then code all 100 responses in terms of the final categories selected. The student was also

given one opportunity to verify the coding system and one opportunity to verify the final codes for the responses. Other details of the simulation are discussed in this chapter and in the appendices.

Figure 5 is used to represent this simulation in the context of the general model for conducting an evaluation effort. Comparing the reliability and validity of the category systems between experimental and control groups constituted Experiment 1, while comparing the reliability and validity of the final codes constituted Experiment 2. Reliability and validity both represent standards of quality for evaluating an evaluation effort, particularly when content analysis methods are used. Creating the category system is a key task for delineating useful information, and coding the responses is a key task for obtaining that information. The purpose of the simulated content analysis was simply to describe and summarize the teachers' judgments about the school district's accountability system. In other words, the student was asked to characterize the union's appraisals of the accountability system. The experimental design associated with this simulation activity is discussed next.

Design of the Study

Two procedurally overlapping experiments were conducted. Both experiments started at the same time but one ended after two tasks while the other ended after four tasks. These four content analysis tasks were to: (1) develop a category system, (2) verify the system, (3) code a set of responses, and (4) verify the codes. Both experiments used a posttest only control group design.

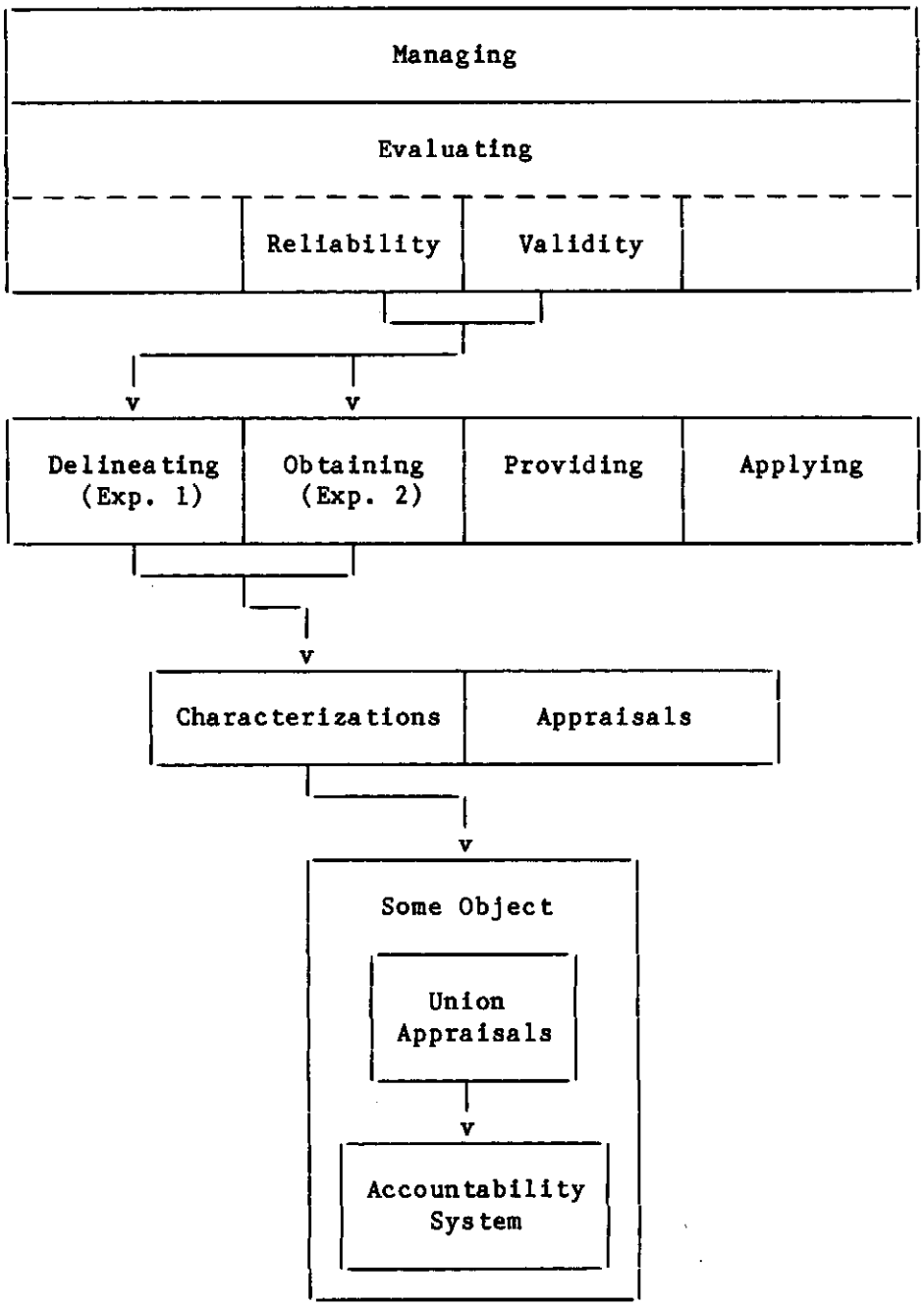


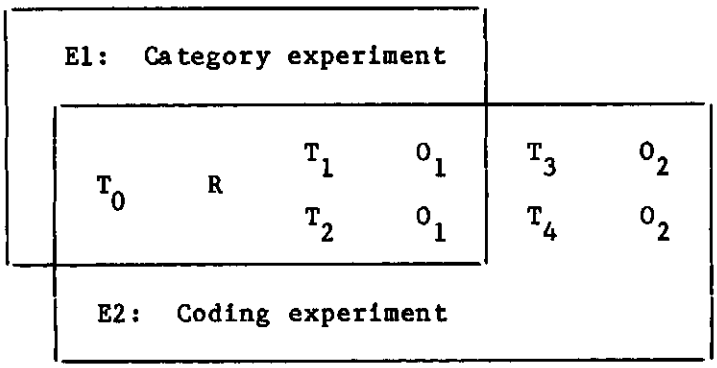
Figure 5. Schematic Representation of the Two Experiments in Terms of the General Model for Conducting an Evaluation Effort

The first experiment was used to test the reliability and validity of a category coding system created with (experimental group) or without (control group) the possession of specially processed computer output. For the first task, each of the experimental participants received a word count list derived from the responses used in the study while the control participants did not. In addition, for the second task, each participant in the experimental group received responses sorted by the codes that the participant used during the previous task. Each group of like-coded responses was headed with the applicable identifier and summary developed by that participant. Each participant in the control group received coded responses in the original order with no identifiers or summaries in that particular document.

The second experiment was used to test the reliability and validity of responses coded into a new set of categories developed by the fictitious research center director. This coding was done with (experimental group) or without (control group) the possession of specially processed computer output. For the third task, participants in the experimental and control groups received materials prepared in basically the same way as for the second task, respectively. The difference was both groups received 50 new responses at the end of their lists and the first 50 response codes were updated to reflect any changes made during the second task. For the fourth task, participants in the experimental group received all 100 responses sorted by the mandatory coding system introduced during the third task. Each group of like-coded responses was headed with the new

identifier and summary used for that group of responses. Again, each participant in the control group only received coded responses in the original order with no identifiers or summaries in the document.

Figure 6 is used to summarize the design. Both experiments started with an in-class training activity (T_0) about relevant theories and practices in content analysis. Participants were then randomly assigned (R) to one of two treatment groups as they received randomly ordered materials for their first task. Experimental participants received special assistance for developing their category systems (T_1) while the control participants did not (T_2). At the end of the category experiment, measures of category reliability and validity for both groups were recorded (O_1). The category experiment procedures stopped here while the coding experiment continued.



Legend. E1 = Experiment 1, E2 = Experiment 2, T_0 = Pre-assignment training, R = Random assignment to groups, T_1 = E1 experimental group treatment, T_2 = E1 control group treatment, O_1 = E1 observation, T_3 = E2 experimental group treatment, T_4 = E2 control group treatment, O_2 = E2 observation.

Figure 6. Design of the Study for Two Procedurally Overlapping Experiments

For the rest of Experiment 2, experimental participants received special assistance coding a set of responses (T_3) while control participants did not (T_4). At the end of the experiment, measures of coding reliability and validity for both groups were recorded (O_2).

Independent Variables

The independent variables for the two experiments of this study were the possession or lack of possession of output from microcomputer programs based on selected content analysis techniques. These outputs included: (a) a word count list sorted by frequency of occurrence derived from the responses used in the study, and (b) the responses sorted and labeled with category identifiers and summaries according to how they were coded by each participant. Experimental participants received these outputs at specific times of the experiments. Control participants received no word count lists and received response lists in the same order before and after they were coded. The sequence in which the materials were distributed and used is discussed in the experimental procedures section. The production of these outputs is discussed in Appendix A and Appendix B.

Research Hypotheses

Four research hypotheses were tested in this study, two each for (a) creating the category system for the content analysis and (b) coding the responses into those categories. One hypothesis for each pair addressed the issue of reliability while the other addressed the issue of validity. The research hypotheses were as follows:

1. Participants who create a category system with the possession of computer output, based on techniques for survey and discovery in content analysis and implemented on microcomputers, will produce more reliable results than participants who create a category system without the possession of such output.

2. Participants who create a category system with the possession of computer output, based on techniques for survey and discovery in content analysis and implemented on microcomputers, will produce more valid results than participants who create a category system without the possession of such output.

3. Participants who code responses with the possession of computer output, based on techniques for survey and discovery in content analysis and implemented on microcomputers, will produce more reliable results than participants who code responses without the possession of such output.

4. Participants who code responses with the possession of computer output, based on techniques for survey and discovery in content analysis and implemented on microcomputers, will produce more valid results than participants who code responses without the possession of such output.

Study Group and Sample

Characteristics of the Study Group and Sample

The study group was comprised of students enrolled in education classes at Western Michigan University during Fall Semester, 1984.

The classes were: ED 322 Teaching of Reading, ED 450/455 School and Society / Educational Perspectives of the Child, ED 516 Symposium on Reading, ED 602 School Curriculum, and EDLD 663 Introduction to Research. The sample for the study consisted of students enrolled in these classes who volunteered to participate in the study.

These classes represent two educational perspectives for the simulation. The first perspective is from a research point of view, EDLD 663. This class emphasizes research methods that can include survey research and content analysis of responses to open-ended survey questions. Thus, participants were given an opportunity to work on a research problem similar to what they might encounter at some point in their careers. The second perspective is from an educational practitioner point of view, the remaining courses. These classes emphasize skills a teacher would use in an elementary or secondary classroom. The survey responses in the simulation were derived from a diverse group of experienced and new teachers, from all curricular areas in a K-12 school system. This encouraged experienced teachers to identify with much of the simulation. It also gave prospective teachers an opportunity to vicariously experience a real-world problem in education.

Sample Size Determination

Because this was considered to be an exploratory study, the minimum sample size needed was determined by using a Type I error level of 0.10, and a ratio of Type I:Type II errors of 1:4 suggested by Cohen "with the idea that the general relative seriousness of

these two kinds of errors is of the order of [1:4] i.e., that Type I errors are of the order of four times as serious as Type II errors" (1977, p. 56). In other words, the probability of rejecting the Null Hypothesis if it were, in fact, true was set at 0.10; and the probability of retaining the Null Hypothesis if it were, in fact, false was set at 0.40. Because power is equal to (1) - (Type II error), the value of power was set at 0.60. In other words, the probability of detecting a statistically significant difference between treatment groups when a difference, in fact, exists was set at 0.60. In addition, the size of the difference between treatment groups needed before a difference could be detected was set at Cohen's suggested medium effect size. "A medium effect size is conceived as one large enough to be visible to the naked eye" (1977, p. 26). Unless a specific rationale to act otherwise exists, a medium effect size is a reasonable choice, especially for an exploratory study.

Cohen's power table (1977, p. 333) for a one-way, fixed-effects analysis of variance was used to determine the minimum number of participants needed to meet the conditions described above. This table was used to determine that at least 28 individuals should participate in each treatment group. Therefore, the total sample size needed to include at least 56 participants.

Random Assignment to Treatment Groups

As a result of the process described below, random assignment of participants to groups was accomplished during the first session when envelopes containing experimental or control group materials were

distributed. First, all participant identification numbers were assigned to one or the other of the two treatment groups. Second, Participant Identification Sheets, Sample of Responses sets, and Category Development Worksheets were labeled and collated by participant identification numbers. The production of these materials are discussed in the next section and in Appendix B. Third, as envelopes were being filled with numbered materials, a Word Count List was added to each envelope with an identification number that had been designated as a experimental participant number. Fourth, the complete packets were arranged in sequential order. Fifth, the packets were systematically passed out to the participants, starting with the lowest available identification number. Packets were passed out systematically by moving in a right-to-left or back-to-front pattern, depending on what was most convenient in relation to the seating pattern.

Procedures

A number of interrelated procedures were implemented for this study. The experiments and the pilot study used to test their procedures are discussed in this section. Three other procedures performed by the researcher and two panels of education and evaluation experts are discussed in this section in terms of how they related to the experimental procedures, but not in great detail. These procedures included the development of a response pool, microcomputer-implemented content analysis activities performed by the researcher for each participant, and the development of a category hierarchy.

Because of the complexity of these procedures and the detail needed to describe them, they are only summarized in this chapter. However, they are also discussed in more detail in the appendices.

Content Analysis Activities Performed by the Researcher and Expert Panelists

The following activities were not performed by the simulation participants, but they still were vital to the success of the experiments. Instead, they were performed by the researcher and two different panels of experts. The activities included: (a) developing a pool of responses to the simulation's open-ended survey question by the researcher and one panel of experts (Appendix A), (b) processing all the individual participants' content analyses by the researcher (Appendix B), and (c) developing a category hierarchy by the researcher and the second panel of experts (Appendix C).

Development of the Response Pool

The response pool was created for two basic purposes. First, it provided the central focus for designing the simulation activity used in both experiments. Second, the category system and codes for the responses also created during the activity provided the basis for judging the validity of the participants' content analyses of the responses. The tasks performed to develop the response pool were to: (a) select a general source of information from which a response pool could be developed, (b) select one open-ended question and its related responses, (c) determine a basic classification framework, (d)

prepare an oversized pool of potential responses, (e) recruit a panel of individuals familiar with the type of information from which the response pool would eventually be created, (f) have the panelists independently create a category system consistent with the basic classification framework selected and code a major portion of the oversized pool of potential responses, (g) process the panelists' work, (h) have the panelists meet to mutually define 10 categories and assign the set of about 200 responses to those categories, (i) select the final five categories and 100 responses, (j) divide the response pool into two groups with roughly equal numbers of responses from each category in each group, and (k) produce a word list for all 100 responses.

Microcomputer-Implemented Content Analysis Activities

Four microcomputer-implemented content analysis activities were conducted by the researcher alone as part of the experimental procedures. These activities were performed after each of the four participant tasks were completed. They were used to: (a) independently process information generated by each participant during a previous task, and (b) prepare individualized materials for each of them to use during the next task, if one followed.

Development of the Category Hierarchy

The category hierarchy was created for two general purposes. First, it was used to derive the measures of category reliability and category validity. Second, it provided a qualitative framework for

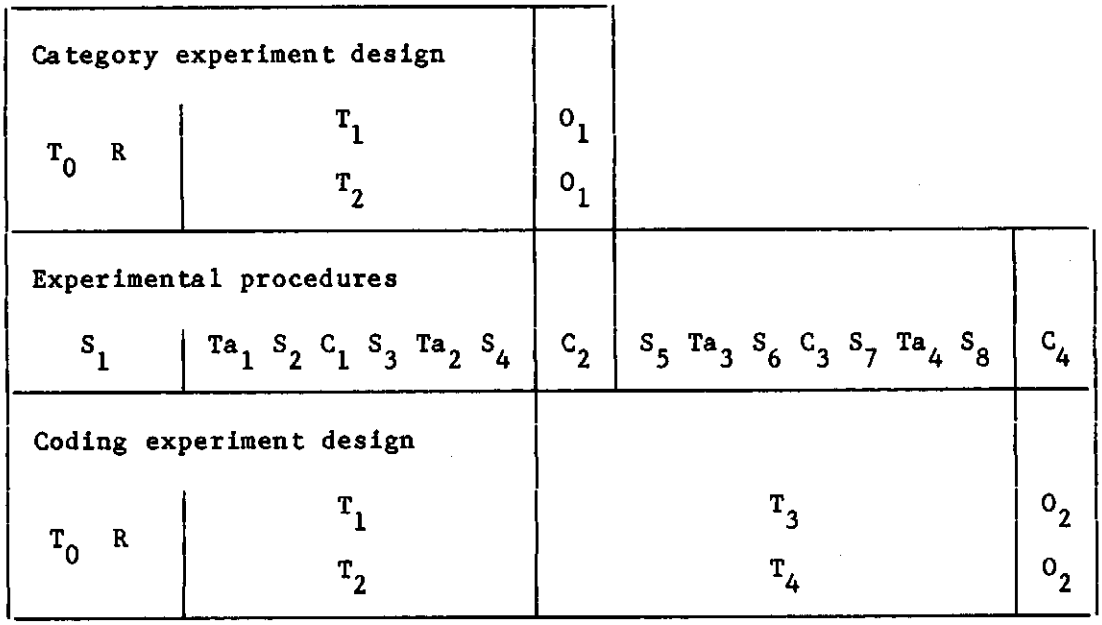
comparing and contrasting how members of the two treatment groups developed their own category systems. The tasks performed to develop the category hierarchy were to: (a) recruit a panel, (b) have the panelists independently create a hierarchy and classify the categories generated by the pilot study participants, (c) process the panelists' work on the pilot study-generated categories, (d) have the panelists cooperatively determine the final framework of the pilot study category hierarchy and assign participant categories to their proper location in the framework, (e) have the panelists independently create a hierarchy and classify the categories generated by the experiment participants, (f) process the panelists' work on the experiment-generated categories, and (g) have the panelists cooperatively determine the final framework of the experiment category hierarchy and assign participant categories to their proper location in the framework.

Experiments

Two procedurally overlapping experiments focused on four content analysis tasks and the activities needed to support the successful completion of those tasks by the participants. The participants in the experiments were students enrolled in education courses at Western Michigan University. These four tasks were to: (1) develop a category system, (2) verify the system, (3) code a set of responses, and (4) verify the codes. The supporting activities included eight group sessions plus four content analysis activities undertaken by the researcher alone. The sessions were designed to provide

classroom instruction, task directions and to exchange materials. The content analysis activities were designed to process the participants' work from a given task and prepare the materials for the next task, if one followed.

Figure 7 is used to summarize the sequential relationships of the classroom sessions, individual tasks, and content analysis activities in the context of the study design. The experiments consisted of eight classroom sessions, four out-of-class tasks for the participants, and four microcomputer content analysis activities conducted by the researcher. One session was conducted each week, and the last session was followed only by the last content analysis activity.



Legend. T_n = Treatment n, R = Random assignment, O_n = Observation n, S_n = Session n, Ta_n = Task n, C_n = Content analysis activity n.

Figure 7. Relationships of the Experimental Designs to the Experimental Procedures

This made the duration of the experimental procedures about eight weeks from start to finish. Both experiments started at Session 1. The category experiment ended with Content Analysis Activity 2. The coding experiment ended with Content Analysis Activity 4. A more detailed description of the tasks and supporting activities follows.

Task 1: Develop a Category System

The first out-of-class task for the participants was to independently develop a content analysis category system. This category system was to be used for coding a set of responses to one open-ended question on a mail survey questionnaire administered by a fictitious university research center to a group of teachers working at a fictitious public school district.

The instructions for completing the task were provided in Read Me First and during the classroom simulation presentation. In brief, the participants were instructed to perform this task independently of all outside help by whatever techniques seemed appropriate and could be completed in one hour or less. They were instructed to write five pairs of category identifiers and summaries. An identifier was a brief, descriptive title for the category that indicated its negative nature. A summary was an operational definition of the category complete enough to allow people besides the student to decide whether a particular response should or should not be included in that category. They were also instructed to put each category number next to each response on the Sample of Responses. Finally, they were instructed to return the Category Development Worksheet and

Sample of Responses to the researcher during Session 2. These items are also discussed in Appendix B.

Group Sessions. The first classroom session was the longest of the eight sessions--about one hour. The three purposes of the session were to: (1) introduce the study, (2) provide a classroom lecture on content analysis, and (3) start the simulation activity that was used as the organizer for the content analysis task to follow. The researcher conducted the session according to a detailed script organized by the three purposes identified above. Using the script ensured a degree of consistency for Session 1 among classrooms. It was not given to any participants.

The participants were given a number of handouts during the session to help them with the first task. In order, they received: (a) Content Analysis: Answers to Four Practical Questions; (b) Read Me First, written instructions for the simulation; (c) a Draft Introduction of an evaluation report; (d) a Practice Exercise; (e) a Practice Exercise Answer Sheet; and (f) the final group of handouts in a sealed envelope. The script and items (a) through (e) are presented in Appendix D. The materials in the sealed handouts are discussed in Appendices A and B. Some envelopes contained materials for experimental participants while others contained materials for control participants. The contents of these envelopes are discussed in the following two sections. A more detailed description of the first session is discussed next.

The introduction was used to provide an advance organizer for the events to follow. It was used to: (a) introduce the researcher,

(b) describe the purpose of the study in general terms, (c) describe the procedures of the study in general terms, (d) notify the students that participation in the study was voluntary, and (e) notify the students that specific incentives for them to complete the study were being offered by the instructor.

One handout was given to the participants during the lecture, Content Analysis: Answers to Four Practical Questions. The information on this handout paralleled the four questions about content analysis addressed in the lecture: (1) What is it?, (2) What are its uses?, (3) When conducting surveys, when should it be used with open-ended questions, instead of using quantitative analysis with forced-choice questions?, and (4) How is it done?

The simulation was started by giving the participants a handout called Read Me First. This handout contained an overview of the simulation and instructions on how to complete the first task. The next handout was a Draft Introduction of an evaluation report being written by Dr. Powerful, the mythical director of the mythical CENTER in the simulation. The Draft Introduction was used to give the participants a richer background of the problem addressed in the simulation. The participants were also given a Practice Exercise and, after completing the exercise, a Practice Exercise Answer Sheet. This practice exercise was intended to give the participants an idea of the types of category identifiers and summaries they should create later on.

The last set of handouts was given to participants in sealed, randomly ordered envelopes. Some packets contained materials for

experimental participants, while others contained control participant materials. A Participant Identification Sheet was attached to the outside of each packet. The identification sheet and handouts in each packet were labeled with the same identification number. Participants filled out the sheets and immediately returned them to the researcher. All packets contained identical copies of a Sample of Responses (see Appendix B). This sample consisted of 50 randomly ordered responses to the open-ended survey question in the simulation. All packets also contained a Category Development Worksheet (see Appendix B). This was included to provide space to write five pairs of category identifiers and summaries.

Session 2 lasted about five minutes. The purpose of this session was for the participants to return the materials used to complete Task 1.

Experimental Conditions. A Word Count List (see Appendix A) was only in the packets of experimental participants. It contained each word that occurred more than once for all 100 responses used in the simulation, not just those 50 responses presented in the first task. The list was ordered by frequency of occurrence of each word.

Control Conditions. The control group did not receive a Word Count List. This was the only difference between the two treatment conditions for the first task.

Task 2: Verify the Category System

The second task for participants was to confirm the work they completed during Task 1. They were told to first read the Category

Development Worksheet and Sample of Responses. After reviewing the information, they were instructed to make any changes in the identifiers, summaries, or codes for particular responses as they saw fit. The task was expected to take less than one hour to complete. They were also instructed to return the materials during Session 4. This task concluded the category experiment for the participants.

Group Sessions. Session 3 lasted about five minutes. The purpose of this session was for the researcher to give the Task 2 materials, a three-document packet, to the participants.

The first document was a personalized (it was addressed to individual participants) form memo attached to the outside of the packet (see Appendix B). It was used to thank them for participating in the first task and give them instructions for the second task.

Two documents were inside each packet. The first was a Category Development Worksheet with all identifiers and summaries written by a given participant typed in. The second was another Sample of Responses document with two code columns instead of one, the first column was for the code given to each response by a given participant during the first task and the second column was for a new code if the participant chose to make a change.

The second document was prepared differently for experimental and control participants. These documents will be described in the following two sections.

Session 4 lasted about five minutes. The purpose of this session was for the participants to return the Task 2 materials to the researcher.

Experimental Conditions. For experimental participants, responses were sorted into five groups, one for each of the categories a given participant developed. The applicable identifier and summary was also placed just before each group of sorted responses (see Appendix B).

Control Conditions. For control participants, the Sample of Responses document was the same as the one in the first task with the exception of the added column and the codes given to the responses were typed in the first column. The responses were in the same order as they were in the first task and the category identifiers and summaries were not put in the document.

Task 3: Code a Set of Responses

The third out-of-class task for participants was to code all 100 responses in relation to the new category system developed by the fictitious Dr. Powerful (the Response Panel). These five categories were labeled A through E in order to reduce confusion with each participant's old categories labeled 1 through 5. They were told their old categories may or may not be like the new categories. The task was expected to take about one hour to complete. They were also instructed to return the materials during Session 6.

Group Sessions. Session 5 lasted about five minutes. The purpose of this session was for the researcher to give the Task 3 materials, a three-document packet, to the participants.

Again, the first document was a personalized memo attached to the outside of the packet (see Appendix B). It was used to thank

them for participating in the second task, inform them that Dr. Powerful had just returned from a distant meeting, and give them instructions for the third task.

Two documents were inside each packet. The first was the Official Categories Summary (see Appendix B) developed by Dr. Powerful. It contained the final set of categories to be used by all participants for the remainder of the simulation.

The second document was the Complete Set of Responses (see Appendix B). This document contained 100 responses to the open-ended question used in the simulation survey--the 50 responses previously used plus 50 new responses. It was prepared differently for experimental and control participants.

Session 6 lasted about five minutes. The purpose of this session was for the participants to return the Task 3 materials to the researcher.

Experimental Conditions. The first 50 responses were prepared in the same way they were prepared for Task 2 except that any changes in response codes made by a given participant were reflected in the new document. This means that the experimental participants had their own category identifier and summary before each group of sorted responses. The second 50 responses were added to the end of the document because they had not yet been coded.

Control Conditions. The control participants received the first 50 responses in the original order, prepared as they were for Task 2 but updated to reflect any changes in response codes. The second 50 responses were added to the end of the document.

Task 4: Verify the Codes

The fourth out-of-class task for the participants was to confirm the work they had completed during Task 3. They were instructed to read the Complete Set of Responses and then make any changes in the codes for particular responses as they saw fit. The task was expected to take less than one hour to complete. They were also instructed to return the materials during Session 8. This task concluded the coding experiment for the participants.

Group Sessions. Session 7 lasted about five minutes. The purpose of this session was for the researcher to give the Task 4 materials, a three-document packet, to the participants.

The first document was a personalized memo attached to the outside of the packet (see Appendix B). It was used to thank them for participating in the simulation and give them instructions for the fourth task.

Two documents were inside each packet. The first was another copy of the Official Categories Summary. It was provided to make sure each participant still had a copy of the categories to be used.

The second document was the Complete Set of Responses updated to reflect the codes entered during the third task. It was prepared differently for experimental and control participants.

Session 8 lasted about five minutes. This session allowed the participants to return the Task 4 materials to the researcher.

Experimental Conditions. The processes used to prepare the document for the experimental group were the same as those used to

prepare the Sample of Responses for Task 2, with two exceptions. First, 100 responses were processed instead of 50 responses. Second, the Official Categories were placed before each group of sorted responses, not the set of categories developed by a given participant.

Control Conditions. The processes used to prepare the document for the control group were the same as those used to prepare the Sample of Responses for Task 2, with one exception. One hundred responses were processed instead of 50 responses.

Pilot Study

The purpose of the pilot study was to test the procedures to be used for the fall experiments so that necessary changes could be made before the experiments began. The pilot study was conducted during the 1984 Summer Session at Western Michigan University. One graduate class, EDLD 663 Introduction to Educational Research, was used.

A different class, ED 601 Foundations of Educational Research, was originally scheduled to be used. However, that instructor decided to drop out of the pilot study after seeing the 200 responses the participants would be asked to process. The basic reason for dropping out was that the instructor thought the task was too complex and time consuming for the students and, as a result, they would actively resist participating in the study. Because the second instructor and the Response Panel expressed similar concerns, the size of the response pool was cut in half as described under Development of the Response Pool in Appendix A.

One other problem was encountered when recruiting the second instructor. Because the researcher was required to sign contractual agreements with all instructors who participated in the experiments (see Appendix E), it was planned to sign a similar contract with the pilot study instructor. However, one section of the contract stipulated that the instructor would provide at least one incentive for the students to participate in the study. Three suggested incentives were to: (1) replace one regular assignment, (2) provide a specified number of bonus points, or (3) award the higher of two grades in "borderline" cases. A blank line was also provided in the contract to enter any fourth incentive. The instructor would not agree to any of the above incentives. Because this was, for all practical purposes, the last opportunity to conduct a pilot study before the fall experiments, the researcher decided to forgo signing a contract with the instructor and to conduct the study without any specified incentives for the participants.

Activities

The two experiments required the participants to perform four out-of-class tasks. They also needed to meet at least briefly with the researcher before and after each task. This required eight group sessions to complete the experiments. Because the course was conducted during a summer session, the class met twice a week over about an eight week period. This made it necessary to conduct the pilot study over a four week period. The participants were asked to perform one task per week. They were given the assignments during the

Tuesday class period and asked to return them during the Thursday class period. The researcher processed their work between Thursdays and Tuesdays. The activities of the pilot study contained all of the procedures described under the Experimental Procedures section plus each member of the class was given a Content Analysis Study Participation Survey during Session 6 to obtain information for improving the procedures.

Results of the Participation Survey

Out of 30 people who attended Session 1, nine people completed the category experiment and eight people completed the coding experiment. Twenty-one people dropped out of the pilot study by Session 2. Thus, the drop-out rate for the category experiment was 70 percent, and the drop-out rate for the coding experiment was 73 percent. Because of the sequential nature of the experiments, three people who were absent for the first session could not participate in the study.

Four problems, as reported by the participants, seem to have contributed to the high drop-out rate in the pilot study. First, there was a shortage of time to complete the tasks. This was mainly due to the Tuesday-Thursday schedule for completing each task and the compressed calendar for the Summer Session. Second, the presentation during Session 1 did not adequately focus on the tasks to be completed. More emphasis on how content analysis was related to the simulation and exactly what was to be done during the simulation was called for. Third, no rewards for completing the simulation were given and no penalties were exacted for dropping out. Fourth, people

who did not attend the first session could not participate in the pilot study.

Modifications to the Procedures

Because of what was learned during the pilot study, five changes were made for the fall experiments. First, the contract was modified. Language was added to the section on Incentives for the Participants to Complete the Study that suggested ways the instructor could verbally encourage participants. In addition, no instructor was allowed to participate in the study without selecting at least one specific incentive and signing the contract. Second, the Script for Session 1 was modified to place more emphasis on using content analysis for responses to open-ended survey questions. It was also modified to place more emphasis on comparing and contrasting how responses to open-ended versus forced-choice survey questions would be analyzed and reported. Third, Content Analysis: Answers to Four Practical Questions was modified to reflect the changes in the script. Fourth, the Participant Identification Sheet was modified to collect more demographic information about the participants. If a high drop-out rate occurred during the fall experiments, this information would be used to compare the drop-out group with those who completed the experiments. Fifth, each of the eight sessions were conducted on the same day of the week for any given course. This allowed the participants one full week to complete each task. It also meant the experiments took about eight weeks to complete instead of the four weeks needed for the pilot study. Because of the

sequential nature of the experiments and the difficulty of scheduling make-up dates for Session 1, no modifications were made to allow absentee students to start the experiments at a later date.

Data Analyses

Measures of Dependent Variables

Four dependent variables were used to investigate the four research hypotheses. Two dependent variables were derived from raw data generated during the category experiment, the work of the Response Panel, and the Hierarchy Panel. Two other dependent variables were derived from raw data generated during the coding experiment and the work of the Response Panel. These variables are described next.

Category Reliability

Category reliability is defined as the extent to which the same set of categories is created from the simulation documents and responses used in the category experiment under varying circumstances, at different locations, by different participants. This was labeled inter-rater agreement, individual reliability, reproducability, and coder reliability in Table 4. This definition of reliability requires that two or more participants must independently create a category system using the same instructions and the same responses. Differences between participants' category systems represent intra-participant inconsistencies and inter-participant disagreements (Krippendorff, 1980, p. 131).

For category reliability, the measure must reflect the extent to which each participant agrees with the rest of his or her treatment group about which five categories should be included in the response classification system. Krippendorff (1980, p. 138) discusses three different agreement coefficients that could be used here: (1) Cohen's (1960) kappa, (2) Scott's (1955) pi, and (3) his own alpha. Although each coefficient is computed a bit differently, both Krippendorff (1980, p. 138) and Cohen (1960, p. 43) show their coefficients to be identical to Scott's pi under certain circumstances. As such, they are all "interpretable as the proportion of agreement after allowance for chance" (Cohen, 1960, p. 43). Unfortunately, all of these coefficients pose problems for use in this study. According to Krippendorff (1980, p. 138), Scott's pi must be used with only two coders and a very large sample of nominal coding units. He also states his alpha is designed to address this problem and be "a generalization to many coders, many kinds of orders (metric) in data, and for any sample size" (p. 138). It is based on a method suggested by Spiegelman, Terwilliger, and Fearing (1953). However, he provides no suggestions for statistically testing the difference between coefficients obtained from two different groups. Cohen suggests a test for comparing obtained kappas from exactly two individuals (1960, p. 44) but not from two different groups.

Because of these problems, no completely satisfactory measure of category reliability was found. As a compromise, a transformed measure was created that could be tested using analysis of variance procedures. This measure was the median proportion of agreement. It

was computed for each participant after each of the five categories developed during Experiment 1 were assigned to one of the 19 Hierarchy categories by the Hierarchy Panel (see Appendix C for details).

The computation was performed with a researcher-written Turbo Pascal (Borland International, 1983) microcomputer program running on a DEC Rainbow. Three basic steps were performed for each participant. The first step was to determine the number of category agreements between the participant and every other participant in his or her treatment group. One category could be counted no more than once, even if another person had two or more categories identical to it. This made five the maximum possible number of agreements with every other participant. Second, the median number of agreements was found for each participant. Third, this median was divided by five. Thus, the measure of category reliability for each participant could range from 0.0, reflecting no agreement with any other treatment group members, to 1.0, reflecting complete agreement with all other treatment group members.

This measure did not explicitly correct for chance agreements. Because of this, the experimental and control means were first tested against mean scores expected by chance using one sample t-tests. The expected value due to chance alone was computed as (1 / the number of categories established by the Hierarchy Panel), or 1 / 19.

Category Validity

Category validity is defined as the extent to which the set of categories created by participants in the category experiment agrees

with the set of categories created by the Response Panel. This is analogous to what Krippendorff calls semantical validity (1980, pp. 159-162), and it constitutes one type of content validity represented in Table 5. Semantical validity is indicated when an analytical procedure produces results that are in substantial agreement with an external criterion procedure involving expert judges who are familiar with the symbolic nature of the material to be analyzed. Krippendorff also calls this type of validity data-oriented. He further contends it "assesses how well a method of analysis represents the information inherent in or associated with available data" (1980, p. 157).

For category validity, the measure must reflect the extent to which each participant agrees with the Response Panel about which five categories should be included in the response classification system. The measure used was the total number of agreements with the Response Panel. It was derived from the same participant data used to derive the measure of category reliability.

The measure of category validity was computed with a researcher-written dBASE II (Ratliff, 1982) microcomputer program running on a DEC Rainbow. For each participant, it counted the total number of categories that agreed with the five Response Panel categories. Each Response Panel category could be matched by all the participant categories no more than once. Thus, the measure of category validity could range from 0, representing no agreement with the Response Panel categories, to 5, representing complete agreement with the Response Panel categories.

Coding Reliability

Coding reliability is defined as the extent to which the same codes are assigned to the responses used in the coding experiment under varying circumstances, at different locations, by different participants. Thus, the discussion of inter-rater agreement under the measure of category reliability also applies here.

For coding reliability, the measure must reflect the extent to which each participant agrees with the rest of his or her treatment group about how the 100 responses should be coded. The measure used was the median proportion of agreement. This time, however, the expected value of the measure due to chance alone is 1 / 5 (1 / the number of categories used by all participants).

The measure for coding reliability was derived for each participant with the same Turbo Pascal program used to derive the measure of category reliability. This was possible because the program was designed to check for which experiment was currently being processed and use the appropriate raw data and equations in the computations. In this case, the three steps were to: (1) determine the number of responses for which the participant assigned codes identical with those assigned by each other treatment group member, (2) find the median number of agreements for each participant, and (3) divide this median by 100. Thus, the measure of coding reliability could range from 0.0, reflecting no agreement with any other treatment group members, to 1.0, reflecting complete agreement with all other treatment group members.

Coding Validity

Coding validity is defined as the extent to which the codes assigned to the responses by participants in the coding experiment agree with the codes assigned to those responses by the Response Panel. Thus, the discussion of the concept of semantical validity under the measure of category validity applies here as well.

For coding validity, the measure must reflect the extent to which each participant agrees with the Response Panel about how the 100 responses should be coded. The measure used was the total number of agreements with the Response Panel.

The measure of coding validity was also computed with a researcher-written dBASE II program running on a DEC Rainbow. For each participant, it counted the total number of responses that agreed with the 100 Response Panel codes. Thus, the measure of coding validity could range from 0, representing no agreement with the Response Panel codes, to 100, representing complete agreement with the Response Panel codes.

Analyses of Dependent Variables

Four research hypotheses were postulated for this study. The reliability hypotheses, Hypotheses 1 and 3, were tested by comparing differences between mean median proportions of agreement between the experimental and control groups. The validity hypotheses, Hypotheses 2 and 4, were tested by comparing differences between mean total scores of agreement with a standard. The null forms of these

hypotheses were tested using one-way analysis of variance procedures for independent samples. One-way analysis of variance was selected because: (a) one independent variable with two levels was used--possession or lack of possession of specialized computer outputs, (b) the groups were independently formed through random assignment, (c) the means of the measures of the dependent variables were considered to be on at least an interval scale, (d) due to random assignment, a normal distribution of scores was assumed, and (e) due to essentially equal numbers of participants in each treatment group, homogeneity of variance was assumed. The analysis of variance tests were computed with researcher-written SPSS (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975) computer programs running on a DECsystem-10.

Summary

The effects of using microcomputer output that was or was not based on content analysis techniques for survey and discovery were examined. Two procedurally overlapping experiments were conducted. The first experiment was used to test the effects of using or not using the specialized microcomputer output on the reliability and validity of developing a category system for a set of responses to an open-ended survey question. The second experiment was used to test the effects of using or not using the specialized microcomputer output on the reliability and validity of coding a set of responses to the same survey question when a category system was already supplied. Participants in the experiments were students enrolled in a number of classes at Western Michigan University.

Both experiments started at the same time, during a classroom lecture on Content Analysis, but the first experiment ended after Task 2 while the second experiment ended after Task 4. The first task for the participants was to create a category system to code a set of responses to an open-ended survey question. The survey was directed toward the teachers working at a fictitious public school district. The question solicited any reactions they had to the school's controversial accountability system. The second task was for participants to verify their work after the researcher differentially processed the materials, based on membership in the experimental or control group. The third task was for participants to code a set of responses to the question based on a category system developed by a fictitious professor conducting the study. The fourth and final task was for participants to verify their work after the researcher differentially processed the materials, based on membership in the experimental or control group.

Development of the response pool, the category system for the fictitious professor, and the information for making judgments about the validity of participants' category systems required the participation of a group of evaluation and education experts called the Response Panel. Deriving the measures of category reliability and category validity also required the participation of a group of evaluation and education experts called the Hierarchy Panel. Differentially processing the participants' work during the course of the experiment required the use of a number of microcomputer-based procedures. A pilot study was also conducted to test the procedures.