The Use of Microcomputer Programs to Improve the Reliability and Validity of Content Analysis in Evaluation

Richard D. Frisbie
Western Michigan University

Follow this and additional works at: http://scholarworks.wmich.edu/dissertations

Part of the Instructional Media Design Commons

Recommended Citation
Frisbie, Richard D., "The Use of Microcomputer Programs to Improve the Reliability and Validity of Content Analysis in Evaluation" (1986). Dissertations. 2294.
http://scholarworks.wmich.edu/dissertations/2294
THE USE OF MICROCOMPUTER PROGRAMS TO
IMPROVE THE RELIABILITY AND VALIDITY
OF CONTENT ANALYSIS IN EVALUATION

by

Richard D. Frisbie

A Dissertation
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Education
Department of Educational Leadership

Western Michigan University
Kalamazoo, Michigan
April 1986
INFORMATION TO USERS

This reproduction was made from a copy of a manuscript sent to us for publication and microfilming. While the most advanced technology has been used to photograph and reproduce this manuscript, the quality of the reproduction is heavily dependent upon the quality of the material submitted. Pages in any manuscript may have indistinct print. In all cases the best available copy has been filmed.

The following explanation of techniques is provided to help clarify notations which may appear on this reproduction.

1. Manuscripts may not always be complete. When it is not possible to obtain missing pages, a note appears to indicate this.

2. When copyrighted materials are removed from the manuscript, a note appears to indicate this.

3. Oversize materials (maps, drawings, and charts) are photographed by sectioning the original, beginning at the upper left hand corner and continuing from left to right in equal sections with small overlaps. Each oversize page is also filmed as one exposure and is available, for an additional charge, as a standard 35mm slide or in black and white paper format.*

4. Most photographs reproduce acceptably on positive microfilm or microfiche but lack clarity on xerographic copies made from the microfilm. For an additional charge, all photographs are available in black and white standard 35mm slide format.*

*For more information about black and white slides or enlarged paper reproductions, please contact the Dissertations Customer Services Department.
Frisbie, Richard D.

THE USE OF MICROCOMPUTER PROGRAMS TO IMPROVE THE RELIABILITY AND VALIDITY OF CONTENT ANALYSIS IN EVALUATION

Western Michigan University

Ed.D. 1986

University Microfilms International 300 N. Zeeb Road, Ann Arbor, MI 48106
PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark √.

1. Glossy photographs or pages ______
2. Colored illustrations, paper or print ______
3. Photographs with dark background ______
4. Illustrations are poor copy ______
5. Pages with black marks, not original copy ______
6. Print shows through as there is text on both sides of page ______
7. Indistinct, broken or small print on several pages ______
8. Print exceeds margin requirements ______
9. Tightly bound copy with print lost in spine ______
10. Computer printout pages with indistinct print ______
11. Page(s) _______ lacking when material received, and not available from school or author.
12. Page(s) _______ seem to be missing in numbering only as text follows.
13. Two pages numbered _______. Text follows.
14. Curling and wrinkled pages ______
15. Dissertation contains pages with print at a slant, filmed as received ______
16. Other _____________________________________________________________

______________________________________________________________

University
Microfilms
International

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Dedicated to the memory of my father, Stanley K. Frisbie,
and to my mother, Norma E. Frisbie.
ACKNOWLEDGMENTS

This study could not have been completed without the hard work and support of a wide variety of individuals. These people provided resources, performed specialized tasks, supervised my work, and gave me the inspiration to continue.

Dan Stufflebeam gave me a place to "temporarily" sit down and think about evaluation, life, and how to play with those new toys—microcomputers. He also made the resources of the Evaluation Center available to me so I could actually perform this study. Because I was fortunate enough to win one of the Harold and Beulah McKee Graduate Research Fellowships, I also was able to more quickly complete the study and disseminate its results to the academic community.

The simulated evaluation project required the participation of several different groups of people. The Response Panel developed a set of responses to an open-ended survey question and created a conceptual framework for analyzing them. These panel members were Mary Anne Bunda, Ron Crowell, Richard Harring, and Tom Ryan. Dave Cowden and his students participated in the pilot study used to test the experimental procedures. Mary Anne Bunda, Joe Chapel, Ron Crowell, Richard Harring and their students participated in the full-scale simulation on which the experiments were based. The Hierarchy Panel analyzed the categories developed by the pilot study and experiment participants, and created a comprehensive framework for classifying all of them. Next to writing the dissertation, this was the
most difficult task of the whole study. The people on this panel were Cynthia Halderson, Dennis Shouse, and Eileen Stryker.

I was very fortunate to assemble a highly talented and knowledgeable doctoral committee. Simply asking myself how they would react to something I was about to do always made me work a little more and a little harder before they found out about it. When they were given the extra advantage of actually seeing something I had written, they always gave me solid advice on how to make it even better. Howard Poole, Jim Sanders, and Dan Stufflebeam all made this a better effort by their presence on my committee. Mary Anne Bunda, my doctoral committee chair, allowed me to follow my feelings or use my head, whenever either seemed appropriate. Not only that, she made sure I finished the dissertation within this lifetime.

Behind all this scholarly activity, my family stood in unwavering support. My puppies, Shanti and Akasha, were always ready to sit on my foot, steal my socks, or go for a good race through the snow. My babies—Amanda, Aleisha, and one to be named at a later date—gave me the best of reasons to be done and to get on with the rest of our lives. My brother and sisters—Stan, Shilah, Jo-Lynn, and Sandie—encouraged me to do well, and they convinced me I could. My parents, Stan and Norma, gave me life. They also gave me the spirit to strive for and become whatever I chose to be. Finally, my soul mate, Linda, gave me her love and showed me the path to a new awakening.

Richard D. Frisbie
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................... ii  
LIST OF TABLES .................................................................................................................. vii  
LIST OF FIGURES ............................................................................................................... ix  

CHAPTER

I. INTRODUCTION ........................................................................................................... 1  
   Need for the Study ........................................................................................................... 1  
   Purpose and Objectives of the Study ............................................................................. 4  

II. REVIEW OF THE LITERATURE ................................................................................. 8  
   Introduction .................................................................................................................. 8  
   Key Concepts for Identifying the Relationships Between Evaluation and Content Analysis .................................................................................................................. 9  
   Key Components of Evaluation and Content Analysis Information ...................... 10  
   Actions of Evaluation and Content Analysis ............................................................ 22  
   Standards of Quality .................................................................................................... 27  
   Working Definitions of Evaluation and Content Analysis ...................................... 36  
   An Overview of Evaluation ......................................................................................... 37  
   Classification of Approaches .................................................................................... 38  
   Summary of Approaches ............................................................................................ 42  
   A General Model for Conducting an Evaluation Effort ........................................... 52  
   An Overview of Content Analysis ............................................................................. 57  
   Classification of Uses ................................................................................................. 57  
   Summary of Uses ....................................................................................................... 59  

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>Table of Contents—Continued</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tasks of Content Analysis ........................................ 74</td>
</tr>
<tr>
<td></td>
<td>Performing Content Analysis Tasks with Microcomputers ......... 87</td>
</tr>
<tr>
<td></td>
<td>Summary ..................................................................... 92</td>
</tr>
<tr>
<td>III. METHODOLOGY</td>
<td>................................................................. 96</td>
</tr>
<tr>
<td></td>
<td>Introduction ................................................................ 96</td>
</tr>
<tr>
<td></td>
<td>Overview of the Study ............................................... 96</td>
</tr>
<tr>
<td></td>
<td>Review of the Problem ................................................ 96</td>
</tr>
<tr>
<td></td>
<td>Simulation Activity ..................................................... 99</td>
</tr>
<tr>
<td></td>
<td>Design of the Study .................................................... 100</td>
</tr>
<tr>
<td></td>
<td>Independent Variables ................................................. 104</td>
</tr>
<tr>
<td></td>
<td>Research Hypotheses .................................................... 104</td>
</tr>
<tr>
<td></td>
<td>Study Group and Sample ............................................... 105</td>
</tr>
<tr>
<td></td>
<td>Procedures ................................................................ 108</td>
</tr>
<tr>
<td></td>
<td>Content Analysis Activities Performed by the Researcher and Expert Panelists ..................................... 109</td>
</tr>
<tr>
<td></td>
<td>Experiments ................................................................ 111</td>
</tr>
<tr>
<td></td>
<td>Pilot Study .................................................................. 121</td>
</tr>
<tr>
<td></td>
<td>Data Analyses ................................................................ 125</td>
</tr>
<tr>
<td></td>
<td>Measures of Dependent Variables .................................... 125</td>
</tr>
<tr>
<td></td>
<td>Analyses of Dependent Variables .................................... 130</td>
</tr>
<tr>
<td></td>
<td>Summary ...................................................................... 131</td>
</tr>
<tr>
<td>IV. RESULTS</td>
<td>................................................................. 133</td>
</tr>
<tr>
<td></td>
<td>Introduction ................................................................ 133</td>
</tr>
<tr>
<td></td>
<td>Effects of Microcomputer Output on Category Development .................................................. 134</td>
</tr>
<tr>
<td></td>
<td>Category Reliability ..................................................... 135</td>
</tr>
</tbody>
</table>
Table of Contents—Continued

CHAPTER

Category Validity .......................... 138

Effects of Microcomputer Output on Response Coding ................. 140

Coding Reliability .......................... 141

Coding Validity ............................ 144

Comparisons of Those Who Did or Did Not Complete Each Experiment 146

Summary ........................................... 152

V. DISCUSSION ................................. 153

Introduction ................................... 153

Interpretation of the Results .................................. 153

Limitations of the Study ..................... 154

Design Limitations ............................ 154

Plausible Explanations for the Retention of the Null Hypotheses of Experiment 1 155

Plausible Explanations for Differential Participation ............... 155

Benefits of the Study ............................ 163

Suggestions for Future Research ................. 165

Closing Comments ............................. 168

APPENDICES

A. Development of the Response Pool .......................... 170

B. Microcomputer-Implemented Content Analysis Activities .......... 184

C. Development of the Category Hierarchy ...................... 195

D. Simulation Materials for Session 1 .......................... 205

E. Researcher/Instructor Contract ............................. 228

BIBLIOGRAPHY .......................................... 233

vi
# LIST OF TABLES

1. Surface and Underlying Contrasts for Evaluation and Content Analysis Information .................................. 15

2. Four Types of Information Derived From Underlying Evaluation and Content Analysis Contrasts .................. 15

3. Six Basic Actions for Conducting an Evaluation or Content Analysis Effort ........................................... 26

4. Classifications of Reliability by Type of Agreement Represented ......................................................... 32

5. Classifications of Validity ............................................. 35

6. Classification of Approaches for Conducting Evaluations Based on Epistemology, Major Perspective, and Orientation .......................................................... 41

7. Summary of Approaches for Conducting Evaluations ................................................................. 44

8. Uses of Content Analysis by Purpose, Communication Element, and Question .................................. 60

9. Group Means and Standard Deviations for Hypothesis 1, Category Reliability ........................................ 137

10. Summary of t-Tests Between Observed Mean Scores and Corresponding Expected Scores Due to Chance Alone for Hypothesis 1, Category Reliability .................................................. 137

11. One-Way Analysis of Variance for Hypothesis 1, Category Reliability ................................................. 138

12. One-Way Analysis of Variance for Hypothesis 2, Category Validity .................................................. 139

13. Group Means and Standard Deviations for Hypothesis 2, Category Validity ........................................ 139

14. Group Means and Standard Deviations for Hypothesis 3, Coding Reliability ........................................ 143

15. Summary of t-Tests Between Observed Mean Scores and Corresponding Expected Scores Due to Chance Alone for Hypothesis 3, Coding Reliability ............................................. 143
List of Tables—Continued

16. One-Way Analysis of Variance for Hypothesis 3, Coding Reliability .................. 144

17. One-Way Analysis of Variance for Hypothesis 4, Coding Validity ......................... 145

18. Group Means and Standard Deviations for Hypothesis 4, Coding Validity ................. 145

19. Characteristics of Students Enrolled in the Six Study Classes ............................... 147

20. Summary of Chi-Square Tests of Differences on Selected Variables Between Those Who Did or Did Not Complete Experiment 1 ........................................... 150

21. Summary of Analysis of Variance Tests of Differences on Selected Variables Between Those Who Did or Did Not Complete Experiment 1 ........................................... 150

22. Summary of Chi-Square Tests of Differences on Selected Variables Between Those Who Did or Did Not Complete Experiment 2 ........................................... 151

23. Summary of Analysis of Variance Tests of Differences on Selected Variables Between Those Who Did or Did Not Complete Experiment 2 ........................................... 151

24. Summary of Analysis of Variance Tests for the Four Study Hypotheses ..................... 152
LIST OF FIGURES

1. The Key Components of Evaluation and Content Analysis Information in Terms of Their Underlying Relationships .......... 19
2. The General Model for Conducting an Evaluation Effort .......... 53
3. Decision Network Using the General Model for Conducting an Evaluation Effort ............................................. 55
4. Uses of Content Analysis in Terms of the Working Definition and the Communication Paradigm .............................. 62
5. Schematic Representation of the Two Experiments in Terms of the General Model for Conducting an Evaluation Effort .......... 101
6. Design of the Study for Two Procedurally Overlapping Experiments ................................................................. 103
7. Relationships of the Experimental Designs to the Experimental Procedures ............................................................ 112
CHAPTER I

INTRODUCTION

Need for the Study

Evaluation practitioners often analyze narrative information in order to determine its relevance to particular study questions. The general approach best suited to this activity is called content analysis, a highly structured and systematic technique.

However, because of the inherent complexities of developing appropriate analytical constructs and the sheer volume of information that must often be analyzed, content analysis is a very difficult process to perform well. Consequently, evaluation practitioners often use impressionistic or overly simplistic methods when a content analysis of narrative information is required. As a result, their analyses often suffer from low reliability and even lower validity. They could do much better analyses by using certain content analysis techniques. The problem is, up until now, the techniques have been very difficult to use.

Fortunately, recent advancements in microcomputer technology have presented new opportunities for effectively and efficiently coping with the inherent complexities of content analysis. First, the growing availability of relatively inexpensive microcomputers has put the power of the computer age on millions of desks in American offices. Second, these machines are usually equipped with an arsenal
of general purpose programs for word processing, data base management, and the like. While many general purpose programs appear to be adaptable to content analysis methods, they were not designed with this use in mind. As a result, they are not accompanied with suggestions for how they might best be used as content analysis tools. To remedy this situation, advice and demonstrations of how practicing evaluators can use general purpose programs to implement accepted content analysis procedures on narrative information are needed.

Narrative information includes documents like books, requests for proposals, plans, and reports; or responses to open-ended questionnaire or interview questions. Structured content analyses of documents are often very difficult, tedious, and time consuming to complete, particularly as the size of the documents increases. Because of this, such analyses are usually performed impressionistically, but often with very useful results. Examples include reviews of books, movies and news conferences. While computers can aid in the analysis of large bodies of narrative information like these, it is not a substitute for the critical, insightful, and sometimes colorful analyses and syntheses done by human experts.

On the other hand, content analyses of responses to open-ended questions are typically performed more systematically—usually with specific questions, variables, and possible responses in mind. Such systematic analyses are much easier to at least partially implement on a computer than the typically ambiguous designs used to analyze large documents. As a result, evaluation practitioners are more likely to immediately benefit from using computerized techniques
focused on content analysis of responses to open-ended questions than from using computerized techniques focused on content analysis of large documents.

Common reasons for using open-ended questions are because the full range of responses to the questions is often not known or a wide range of responses is expected. Such questions also allow people to put the responses in their own words rather than be forced to choose from a fixed list of responses. In addition, open-ended evaluation questions are usually tailored to the unique characteristics of a particular study. As a result, new analytical constructs for describing or interpreting the responses must often be developed. In practice, analytical constructs are operationalized by way of the category systems into which the narrative information is coded. If computerized techniques can be used to help practicing evaluators develop these category systems, the reliability and validity of the studies in which they employ the techniques should improve immensely.

Furthermore, survey questions are usually asked of a large number of people. When these questions are open-ended, the volume of narrative responses that must be analyzed can become quite large. As a result, computerized techniques to help code this large volume of responses more reliably and validly should be a boon to practicing evaluators as well.

In summary, a new opportunity exists for practicing evaluators to use powerful, microcomputer-implemented content analysis techniques in their work. This opportunity results from the following conditions:
1. Evaluators often collect narrative information that is best understood through the use of content analysis.

2. Because of the difficulties in obtaining reliable and valid results from content analyses, evaluation practitioners can benefit from using procedural techniques developed by content analysis experts.

3. Because general purpose microcomputer programs are currently available that can be adapted to content analysis uses, evaluation practitioners can now use computerized techniques that were previously available only to a limited number of researchers with access to very expensive and specialized resources.

4. Because evaluation practitioners are already accustomed to analyzing responses to open-ended questions systematically, they are likely to immediately benefit from using computerized techniques focused on this type of content analysis problem.

5. Two potential areas for improving the reliability and validity of content analyses include the use of computerized techniques to help develop category systems and code large volumes of data.

Purpose and Objectives of the Study

In order to address the needs and opportunities listed above, the purpose of this study is to advance the body of knowledge about how evaluation practitioners can use microcomputer programs to
improve the reliability and validity of content analyses of responses to open-ended survey questions. This purpose can be met by accomplishing the following four objectives:

1. Describe conceptual and operational relationships between evaluation, content analysis, and microcomputers.

Because the fields of evaluation and content analysis evolved from a common heritage of scientific inquiry, they contain many operational similarities. On the other hand, fundamental differences between the conceptual orientations of each field also exist. A description of key evaluation and content analysis concepts allows for the fundamental similarities and differences between them to be identified. These concepts can then be used as organizers for presenting (a) overviews of evaluation and content analysis, (b) a general model for conducting an evaluation effort compatible with content analysis methods, and (c) how microcomputers can be used to help analyze responses to open-ended survey questions. These conceptual and operational relationships are discussed in Chapter 2.

2. Determine the effects of using specialized output from microcomputer programs on the reliability and validity of developing a category system for a large set of responses to an open-ended survey question used in a simulated evaluation effort.

3. Determine the effects of using specialized output from microcomputer programs on the reliability and validity of coding the set of responses in terms of the final category system used.

This dissertation study includes two procedurally overlapping experiments. The focus of the first experiment is on improving the
reliability and validity of developing a content analysis category system. The focus of the second experiment is on improving the reliability and validity of coding the set of responses in terms of an established category system. The setting for the experiments is a simulated evaluation effort that includes analyzing responses to an open-ended survey question. The mock survey was used in an evaluation of a controversial accountability system for a medium sized public school district. Participants in the experiments—volunteer students enrolled in a number of College of Education classes—were first asked to develop a category system for a set of responses to the open-ended question. Midway through the simulation and after the official category system was adopted by the hypothetical project director, the participants were asked to code all of the collected responses in terms of the final categories. The basic methodology used in the experiments is presented in Chapter 3. The results of the experiments are summarized in Chapter 4.

Three complex content analysis activities presented briefly in Chapter 3 are discussed in detail in the appendices. These activities were not performed by the experimental 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).

The accounts contained in the appendices offer researchers interested in conducting studies similar to this one enough detail to adapt the activities to their own purposes. Those readers only interested in the techniques used to help analyze responses to open-ended survey questions can forego the more complex discussions found in Appendices A, B, and C.

4. Identify the limitations and benefits of the study and suggest a program of future research based on the relationships between evaluation, content analysis, and microcomputers; and the results of the study.

The use of microcomputers for content analyses of responses to open-ended survey questions has the potential to make an important contribution to evaluation practice. Chapter 5 is used to (a) present an interpretation of the research results, (b) identify the limitations of the study, (c) identify the benefits of the study, and (d) propose a program of future research based on the relationships between evaluation, content analysis, and microcomputers; and the results of the study.
CHAPTER II

REVIEW OF THE LITERATURE

Introduction

The opportunity for evaluation practitioners to use microcomputer programs when conducting content analyses of responses to open-ended survey questions now exists. However, the best use of this opportunity requires a sound understanding of the conceptual and operational relationships between evaluation, content analysis, and microcomputers. This chapter is intended to promote such an understanding so that evaluation practitioners can better address applied content analysis problems used in their work. It is also used to provide the basis for the experimental design elements of this study.

The following discussions are organized into four main topics.

The first topic focuses on key concepts for identifying the relationships between evaluation and content analysis. These concepts are discussed in terms of different types of information, actions, and standards of quality. Working definitions of evaluation and content analysis that use these concepts are then presented.

The second topic focuses on an overview of evaluation at two levels of detail. First, classification systems that have been used to describe alternative approaches to conducting evaluations are presented and consolidated into one framework. Second, each of the alternative approaches is presented in further detail.
analysis is placed into this framework as one type of evaluation approach.

Third, a general model for conducting an evaluation effort is presented. This model is based on the working definition of evaluation. As such, it is general enough to accommodate content analysis as part of an evaluation effort.

Fourth, an overview of content analysis focuses on its uses, tasks, and microcomputer implementations. In the first two discussions, several uses of content analysis are classified and then summarized. In the summary, those uses particularly relevant to evaluation are highlighted. Then, tasks crucial to obtaining high quality content analysis information are discussed. Finally, the ways microcomputers can help people perform many of these tasks are presented. These operational relationships provide the basis for designing focused studies on how microcomputers can be used to help people perform more reliable and valid content analyses used in evaluation efforts.

Key Concepts for Identifying the Relationships Between Evaluation and Content Analysis

Evaluation is often thought of as the process of describing and judging some object (e.g., Guba & Lincoln, 1981; Joint Committee, 1981; Stake, 1967; Worthen & Sanders, 1973), while content analysis is often thought of as the process of describing and making inferences about some object (e.g., Holsti, 1969; Osgood, 1959; Stone, Dunphy, Smith & Ogilvie, 1966). Even though most authors use verb
forms of these concepts to represent actions, they are better thought of in their noun forms for this study—descriptions, judgments, and inferences—as types of information. As such, both of these enterprises have in common the process of developing a body of information about some object. Nevertheless, the basic actions people perform in order to develop evaluation or content analysis information also turn out to be quite similar (e.g., Krippendorff, 1980; Stufflebeam et al., 1971). The main difference between evaluation and content analysis is based on the underlying contrasts used to partition the information. These different underlying contrasts lead to different connotations for common terms, and different standards for judging the quality of practice. Because these are important issues, this section is used to (a) identify the key components of evaluation and content analysis information in terms of their underlying relationships, (b) identify basic actions used to develop evaluation and content analysis information, (c) discuss different standards of quality that have emerged for judging the information and related processes, and (d) present working definitions of evaluation and content analysis.

Key Components of Evaluation and Content Analysis Information

Even though evaluation can be used to develop descriptions and judgments about an object while content analysis can be used to develop descriptions and inferences about it, this partitioning is not as clear cut as it first appears. The term, descriptions, does not have quite the same meaning to evaluation theorists as it does to
content analysis theorists. In addition, the terms, judgments and inferences, are more similar in meaning to evaluation and content analysis theorists than one might first expect. Fortunately, the conceptual similarities and differences between the various types of information can be clarified by examining the underlying conceptual contrasts and identifying other terms that more clearly reflect these contrasts. This is accomplished by identifying the applicable underlying contrasts, combining them to clarify their relationships, and redefining the key types of evaluation and content analysis information based on these relationships. The main benefit from this examination is a more refined perspective on the fundamental characteristics of both evaluation and content analysis information.

Evaluation Contrasts

Webster's dictionary (Gove, 1971, p. 786) defines the term, evaluate, as "to examine and judge concerning the worth, quality, significance, amount, degree or condition of." The American Heritage dictionary (Morris, 1969, p. 453) defines the term, evaluate, as "1. To ascertain or fix the value or worth of. 2. To examine and judge; appraise." These definitions support two basic premises for this study. First, evaluation is values-based. This premise is important because it implies all true evaluation efforts must be values-based and any efforts that are not values-based are not evaluation efforts. They are something else. This premise is also consistent with Scriven's contention that evaluation "goals always include the estimation of merit, worth, [or] value" (1967, p. 42).
The second premise is that evaluation involves two fundamental information components. For now, the terms descriptions and judgments will be used to represent these components. Many definitions of evaluation include these two components, although at varying levels of specificity. The dictionaries refer to these components as examination and judgment. Stake (1967) claims, "Both description and judgment are essential—in fact, they are the two basic acts of evaluation" (p. 109). Worthen and Sanders (1973) state, "Evaluation is the determination of the worth of a thing. It includes obtaining information for use in judging the worth of [some object]" (p. 19). Guba and Lincoln (1981) "define evaluation as a process for describing an evaluand and judging its merit or worth" (p. 35). Finally, the Joint Committee on Standards for Evaluations of Educational Programs, Projects, and Materials (Joint Committee, 1981) defines evaluation as "the systematic investigation of the worth or merit of some object" (p. 12).

Thus, the typical or surface contrast made between two fundamental components of evaluation information is that of descriptions vs. judgments. This contrast is based on an underlying contrast that has its roots in philosophy. Epistemology is the philosophy of knowledge. It addresses questions about how we can discover and know the truth or facts about some object. Ethics is the philosophy of values. From ethics we determine what is good or bad, what is right or wrong, and what we ought to do in a given situation—rules of conduct. Judgments are ultimately based on ethical principles—value statements.
Content Analysis Contrasts

Berelson (1952) provides the classic definition of content analysis. He states, "Content analysis is a research technique for the objective, systematic, and quantitative description of the manifest content of communication" (p. 18). This descriptive orientation represents the classical content analysis approach. However, many theorists consider it to be too narrow.

A broader perspective of content analysis goes beyond description to also include making inferences. For example, Osgood (1959) defines "content analysis as a procedure whereby one makes inferences about sources and receivers from evidence in the messages they exchange" (p. 36). A similar definition was developed jointly by Stone and Holsti, although they published it separately. They contend, "Content analysis is any research technique for making inferences by systematically and objectively identifying specified characteristics within text" (Holsti, 1969, p. 14; Stone et al., 1966, p. 5).

Thus, content analysis is often thought of as involving the process of describing some object and making inferences related to it. The surface contrast is between descriptions vs. inferences. The underlying contrast is based on the means through which information about some object is acquired. Descriptions are based on sensory input from the outside world. Regardless of the actual senses involved (sight, hearing, touch, taste, or smell) or the actual perceiver (human or otherwise), all sensory input can be called observations. Inferences derive from applying the rules of logic to
a set of statements. These statements usually include descriptions based on observations, and other statements—conclusions—based on theoretical principles. The theories and their principles can be grounded in either epistemology or ethics.

The Contrasts Combined

Table 1 is used to summarize the surface contrasts and underlying contrasts for evaluation and content analysis information used in this study. The surface contrast is descriptions vs. judgments for evaluation, and descriptions vs. inferences for content analysis. The underlying contrast is knowledge vs. value statements for evaluation, and observations vs. logic for content analysis.

When the underlying contrasts are crossed in a two-by-two table, four types of information can be identified. This is represented in Table 2. The four types of information include observational knowledge, logical knowledge, observational value statements, and logical value statements.

Observational knowledge is typified by what we commonly call "the facts"—what we can learn simply by observing something. "The union has 197 members," represents a statement of observational knowledge.

Logical knowledge is derived from applying rules of logic to available information and knowledge-seeking, theoretical principles. "The experimental treatment effect was significantly greater than the control treatment effect," represents a statement of logical knowledge.
Table 1
Surface and Underlying Contrasts for Evaluation and Content Analysis Information

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Evaluation</th>
<th>Content analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface</td>
<td>Descriptions vs. Judgments</td>
<td>Descriptions vs. Inferences</td>
</tr>
<tr>
<td>Underlying</td>
<td>Knowledge vs. Value statements</td>
<td>Observations vs. Logic</td>
</tr>
</tbody>
</table>

Table 2
Four Types of Information Derived From Underlying Evaluation and Content Analysis Contrasts

<table>
<thead>
<tr>
<th>Content analysis contrast</th>
<th>Evaluation contrast</th>
<th>Knowledge vs. Value statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations vs. Logic</td>
<td>Observational knowledge</td>
<td>Observational value statements</td>
</tr>
<tr>
<td></td>
<td>Logical knowledge</td>
<td>Logical value statements</td>
</tr>
</tbody>
</table>
Observational value statements of good/bad, right/wrong, or how we should act in a particular situation, are based on the assumption we can observe the value of something or someone in the same way we can observe many of its other attributes, such as color or language spoken. Such an assumption is based on the philosophical doctrine called ethical naturalism (Harrison, 1967). "According to ethical naturalism, moral judgments just state a special subclass of facts about the natural world" (Vol. 3, p. 69). This doctrine has been rejected by G. E. Moore. Harrison, (1967) represents Moore's position as follows:

Moore contended that goodness was a unique, unanalyzable, nonnatural property (as opposed to natural properties, such as yellowness or anger, that are perceived through the senses or through introspection). Therefore, any attempt to define goodness in terms of any natural property must be a mistake that is one form of what he called the "naturalistic fallacy." (Vol. 3, p. 69)

Thus, observational value statements are examples of the naturalistic fallacy. Because of Moore's criticism and the availability of a commonly accepted alternative to be discussed next, observational value statements are not included as an acceptable type of information for this study.

Logical value statements are derived from applying rules of logic to available information and ethical principles. The naturalistic fallacy is not at issue in this situation because no contention is made that these statements represent factual attributes about the object in question. Instead, the statements derive their value components from the specific ethical principles involved.
Evaluators typically distinguish between two types of value statements (e.g., Guba & Lincoln, 1981; Joint Committee, 1981; Scriven, 1967) most often referred to as merit and worth. Guba and Lincoln (1981) use the term, merit, to mean "intrinsic, context-free value" (p. 39). They further state an entity has merit if it has "value of its own, implicit, inherent, independent of any possible applications" (p. 39). When an entity has value within some context of use or application, they use the term, worth. They define it to mean "extrinsic or context-determined value" (p. 40). They acknowledge their terms, merit and worth, are types of value and they are analogous to Scriven's (1978) terms, merit and value; but they claim the use of their terms avoids "the redundancy and confusion that result when one of the subtypes is called by the same name as the more general type" (p. 40). They also acknowledge Scriven's (1967) notions of intrinsic and payoff evaluation and Tyler's (1949) concern for internal checkpoints and desired outcomes allude to the distinctions they make between merit and worth. However, they contend they "have addressed the issue in a more systematic way than Tyler and other earlier writers" (p. 40).

The concept of merit sounds suspiciously like observational value statements, but this need not be the case. If merit is taken to mean the factual value component of some object, then it does represent the naturalistic fallacy and it is not an acceptable type of information for this study. On the other hand, if merit is taken to mean value implicitly generalized to become free of any specific context, then the naturalistic fallacy is avoided. From this
perspective, merit is not a factual attribute of an object but a
generalized depiction of value for that object in relation to a class
of contexts. For example, to say a university professor with several
refereed publications in his or her field has merit, should not be
taken to mean publications are a value-attributed of university
professors. Instead, it should mean, generally speaking, professors
with many refereed publications are of value (i.e., competent and
productive). In this way, the generalized value statement can only
be derived by combining information about the object and a class of
contexts with ethical principles by using rules of logic.

Because this process is often performed implicitly, it can take
on the appearance of an observational value statement. On closer
examination, however, the ethical perspectives necessary to make
determinations of merit can usually be extracted. To avoid the
controversy associated with the naturalistic fallacy, the informa-
tion, ethical principles, and logical transformations used to make
determinations of merit should be explicitly stated.

The concept of worth represents context-specific value state-
ments. As such, the naturalistic fallacy is not at issue because the
value statements are clearly dependent on variable situations that
include different and often conflicting ethical principles. This
makes it impossible for them to be inherent attributes of an object.

The Key Components and Their Underlying Relationships

It is now possible to reconstruct the components of evaluation
and content analysis information that distinguish them from each
other in terms of the underlying conceptual relationships involved. This reconstruction is represented in Figure 1. The terms in boxes represent key components, while the terms spanned by arrows represent underlying concepts. The concept of high quality information has been added to represent the implicit, common feature of all the components and other concepts. Standards of quality for each field will be the focus of a later discussion.

<table>
<thead>
<tr>
<th>Characterizations</th>
<th>Appraisals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptions</td>
<td>Conclusions</td>
</tr>
<tr>
<td>Descriptions</td>
<td>Impartial conclusions</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>High quality information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge</td>
</tr>
<tr>
<td>Generalized</td>
</tr>
</tbody>
</table>

| Observations | Logic |

Figure 1. The Key Components of Evaluation and Content Analysis Information in Terms of Their Underlying Relationships

All concepts directly under a component help define the nature of that component. The terms for the components have been selected to represent the spirit of the applicable underlying concepts and complement the terms for the other components on all three levels. Also, the underlying concepts are arranged so that observations and
value statements do not overlap. This precludes any components from representing the naturalistic fallacy—considering value to be an inherent characteristic of an object.

The first level represents the two basic components of evaluation information. For the remainder of this study they will be called characterizations and appraisals. Characterizations are based on knowledge while appraisals are based on value statements.

The second level represents the two basic components of content analysis information. They will be called descriptions and conclusions. Descriptions are based on observations while conclusions are based on logic.

The third level represents four important subcomponents that are parts of both evaluation and content analysis information. These subcomponents include descriptions, impartial conclusions, determinations of merit, and determinations of worth. They are derived by simultaneously considering all applicable underlying concepts. Descriptions are based on observational knowledge. This subcomponent has the same name as a basic content analysis component because observational value statements have been excluded from consideration. Impartial conclusions are based on logical knowledge. Determinations of merit are based on logical, generalized value statements. Determinations of worth are based on logical, context-specific value statements.

All four subcomponents represent both evaluation and content analysis information, but they are grouped differently under the main components. For evaluation, characterizations are composed of
descriptions and impartial conclusions; while appraisals are composed of determinations of merit and determinations of worth. For content analysis, descriptions are not further subdivided; while conclusions are composed of impartial conclusions, determinations of merit, and determinations of worth.

Summary

The key components of evaluation and content analysis information, and the basis for their relationship have now been identified. The evaluation components, characterizations and appraisals, and the content analysis components, descriptions and conclusions, are based on different underlying contrasts. For evaluation, the key contrast is based on two fundamental branches of philosophy. Characterizations are based on epistemology, the philosophy of knowledge, while appraisals are based on ethics, the philosophy of determining good or bad, right or wrong, and rules of conduct. In modern usage, responses to ethical questions are often called statements of values (Frankenna, 1967). For content analysis, the key contrast is based on two modes of acquiring information about something. Descriptions are based on sensory input, observations. Conclusions are based on rules of logic that incorporate available information and theoretical principles intended to either acquire knowledge or derive value statements.

Four subcomponents of both evaluation and content analysis information have been identified by simultaneously considering the underlying contrasts of each field. One concept, observational value
statements, was excluded from further consideration because it represented the naturalistic fallacy—considering value to be a natural, inherent attribute of an object. It was replaced by further subdividing logical value statements into generalized and context-specific groups. All these components and underlying concepts are considered to represent high quality information. The actions through which this information is developed are the focus of the next section.

**Actions of Evaluation and Content Analysis**

The purpose of this section is to identify some very basic actions that can apply to different evaluation and content analysis approaches. These actions represent the operations evaluation and content analysis have in common. They also represent the action components of a general model to be discussed in a later section. Five sources, four from evaluation and one from content analysis literature, have been used to identify these actions.

**Evaluation Actions**

Stufflebeam et al. (1971) provide a good actions-oriented definition of evaluation. It is defined as "the (process) of [1] (defining), [2] (obtaining), and [3] (providing) (useful) (information) for [4] (judging) (decision alternatives)" (p. 40). Each of the terms in parentheses are further defined by the authors.

The Joint Committee on Standards for Educational Evaluation (Joint Committee, 1981) has a set of thirty standards to be used for the action of evaluating evaluation. These standards are grouped in
a functional table of contents in terms of ten other actions: 

1. Administering Evaluation  
2. Analyzing Information  
3. Budgeting Evaluation  
4. Collecting Information  
5. Contracting Evaluation  
6. Deciding Whether to Evaluate  
7. Defining the Evaluation Problem  
8. Designing Evaluation  
9. Reporting Evaluation  
10. Staffing Evaluation

(pp. xvii-xx). Specific standards can apply to more than one action.

The Evaluation Research Society (ERS) has developed a set of fifty-five standards to be used for the action of evaluating evaluation (ERS Standards Committee, 1982). These standards are divided into six actions-oriented sections: 

1. Formulation and Negotiation,  
2. Structure and Design,  
3. Data Collection and Preparation,  
4. Data Analysis and Interpretation,  
5. Communication and Disclosure, and  
6. Utilization

(p. 11). Each individual standard applies to only one of the above actions.

Brinkerhoff, Brethower, Hluchyj, and Nowakowski (1983) organize the chapters of their book, Program Evaluation: A Practitioner's Guide for Trainers and Educators, by seven evaluation functions. They call these functions: 

1. Focusing the Evaluation  
2. Designing Evaluation  
3. Collecting Information  
4. Analyzing Information  
5. Reporting Information  
6. Managing Evaluation  
7. Evaluating Evaluation (Meta-Evaluation)

(p. v).

Content Analysis Actions

No formal classification system for content analysis actions like those for evaluation exists. However, one actions-oriented
classification system for content analysis activities comparable to those for evaluation was identified. Krippendorff (1980) uses this classification system as the basis for a practical guide to conducting a content analysis. He contends that "any content analysis involves three logically separate activities: [1] design, [2] execution, [3] report" (p. 169).

Consolidation and Summary

The above sources identify several actions that apply to both evaluation and content analysis efforts. However, many of the actions are too specific for what is needed here. When these actions are placed into more general groups, six basic actions of evaluation or content analysis efforts are suggested. The more specific actions are not intended to represent an exhaustive list of components that comprise the basic actions. The groupings have been used only to help identify those basic actions.

Due to size considerations, a comprehensive summary of evaluation tasks that can be subsumed under these six basic actions is not presented. However, the interested reader can find classifications of evaluation tasks from a number of other sources (e.g., Anderson, Soptick, Rogers, & Worthen, 1971; Schalock & Sell, 1972; Stufflebeam, 1973; Worthen & Sanders, 1984). A simple classification of content analysis tasks relevant to the experimental study is discussed in this chapter as part of the overview of content analysis.

Four of the actions focus on processing some kind of information while two focus on the effort itself. These six actions, their
relationships to the literature sources, and their main focus are summarized in Table 3.

First, the four actions of delineating, obtaining, providing, and applying information are suggested. (For evaluation, the two main types of information are characterizations and appraisals. For content analysis, the two main types of information are descriptions and conclusions.) Delineating information involves the general action of specifying what information is needed and how it will be acquired. Delineating encompasses more specialized actions identified in several sources. These actions were called contracting, deciding, defining, designing, formulating, focusing, negotiating, and structuring. Obtaining information involves the general action of acquiring it in its "raw" state and transforming it to a usable state. Obtaining is roughly synonymous with execution. It also encompasses more specialized actions. They were called analyzing, collecting, interpreting and preparing. Providing information involves the general action of delivering it to the appropriate audiences. Providing is similar to communicating, disclosing, and reporting. Applying information involves the general action of using it for intended or unintended purposes. Applying is an action at a level of specificity comparable with the three general actions above. It is roughly synonymous with utilizing. Judging decision alternatives is a specific example of applying (evaluation) information.

Second, the two actions of managing and evaluating an effort are suggested. These actions can be applied to both evaluation and content analysis efforts, even though Krippendorff does not directly
### Table 3

**Six Basic Actions for Conducting an Evaluation or Content Analysis Effort**

<table>
<thead>
<tr>
<th>Literature Source</th>
<th>S</th>
<th>J</th>
<th>E</th>
<th>B</th>
<th>K</th>
<th>Action</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>5, 6</td>
<td>1, 2</td>
<td>1, 2</td>
<td>1</td>
<td>* Delineating</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>7, 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2, 4</td>
<td>3, 4</td>
<td>3, 4</td>
<td>2</td>
<td>* Obtaining</td>
<td>Information</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>9</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>* Providing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>* Applying</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1, 3, 10</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>* Managing</td>
<td>The effort</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>All</td>
<td>7</td>
<td></td>
<td></td>
<td>* Evaluating</td>
<td></td>
</tr>
</tbody>
</table>

mention them. Managing an effort involves the general action of ensuring all required functions are performed appropriately. It is roughly synonymous with administering, and it encompasses the more specific actions of budgeting and staffing. Evaluating an effort involves the general action of characterizing and appraising it. All of the actions mentioned by the Joint Committee and the Evaluation Research Society were used in the context of evaluating an effort. Meta-evaluation is another term that can be used for evaluating evaluation efforts. The actions of interest for evaluating an effort were discussed in this section. Standards of quality for actually judging them are discussed in the next section.

Standards of Quality

Before evaluation or content analysis practice can be improved, some basis for determining what constitutes an improvement must exist. Standards of quality serve this role. Such standards are currently available for evaluation and content analysis practitioners but at different levels of formality. Evaluation practitioners have available to them published standards developed by professional and regulatory sources, while content analysis practitioners do not. Instead, they must rely on informal sources for indicators of quality. Some of these sources for each field are discussed next.

Evaluation Standards

Two sources of the most comprehensive standards for evaluation quality have already been presented. They are the Joint Committee on
Standards for Educational Evaluation (Joint Committee, 1981) and the Evaluation Research Society (ERS Standards Committee, 1982).

**Joint Committee.** The thirty Standards for Evaluations for Educational Programs, Projects, and Materials (Joint Committee, 1981) are presented in four groups that correspond to four main concerns about any evaluation—utility, feasibility, propriety, and accuracy. Each standard is explained and clarified through a commentary which includes an overview of intent, guidelines for application, common pitfalls, caveats (or warnings against being overzealous in implementing the standard), and an illustration of the standard's application. (pp. 1-2)

Eight "Utility Standards are intended to ensure that an evaluation will serve the practical information needs of given audiences" (p. 19). Three "Feasibility Standards are intended to ensure that an evaluation will be realistic, prudent, diplomatic, and frugal" (p. 51). Eight "Propriety Standards are intended to ensure that an evaluation will be conducted ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results" (p. 63). Finally, eleven "Accuracy Standards are intended to ensure that an evaluation will reveal and convey technically adequate results about the features being studied that determine its merit or worth" (p. 97).

**Evaluation Research Society.** The fifty-five Evaluation Research Society Standards for Program Evaluation (ERS Standards Committee, 1982) are divided into six sections. The sections are listed in roughly sequential order for an evaluation effort. The individual standards are presented as "simple admonitory statements" (p. 11).
Twelve standards of Formulation and Negotiation are based on the assumption, "before an evaluation program or project is undertaken, the concerned parties should strive for a clear mutual understanding of what is to be done, how it is to be done and why, and for an appreciation of possible constraints or impediments" (p. 12).

Six Structure and Design standards are presented because "the design for any evaluation cannot be conceived in a vacuum. It is necessarily influenced by logistical, ethical, political, and fiscal concerns and therefore must take these as well as methodological requirements into account" (p. 13).

The section on Data Collection and Preparation includes twelve standards. These standards are based on the assumption a sound design and work plan have been developed. However, circumstances can change and these activities might need to be altered to reflect those changes.

Nine standards address Data Analysis and Interpretation. They are also based on the assumption a sound design and work plan have been developed, but analyses must be tempered to reflect the data actually collected.

Ten standards address Communication and Disclosure. Good communication is important in order to clarify the nature of the program, the expectations for the evaluation, and even the type of evaluation required . . . ; to anticipate restrictions on release of results and potential conflicts of interest . . . ; to establish accountability for the effort . . . ; to secure the cooperation of parties involved in the program and the evaluation . . . ; and to distinguish objective findings clearly from opinion and interpretation. (pp. 15-16)
Finally, six standards address the Use of Results. These standards are based on the assumption, "the use of evaluation results cannot be guaranteed, of course, but it will be more likely if careful attention is given to the information needs of potential users of the results throughout all phases of the evaluation" (p. 16).

Comparisons have also been made between the Evaluation Research Society and the Joint Committee Standards (e.g., Braskamp & Mayberry, 1982; Cordray, 1982; Stufflebeam, 1982). A common conclusion has been the similarities in the issues addressed and the expectations for quality, by and large, outweigh the differences. This suggests a high degree of agreement about the standards of quality exists in the field of evaluation.

Other Standards. Other evaluation standards have been written for more specialized purposes or audiences. For example, the U.S. General Accounting Office (1978) has a set of standards for assessing social program impact evaluations; and the U.S. Department of Education (1981) has published criteria to help select funding proposals submitted to the Office of Special Education that have sound evaluation designs.

Content Analysis Standards

No published standards for content analysis practice comparable to those for evaluation practice exist. However, two criteria for judging the quality of content analysis efforts—reliability and validity—are mentioned by a number of authors (e.g., Andrén, 1981;
Berelson, 1952; Budd, Thorp, & Donohew, 1967; Carney, 1972; Holsti, 1969; Janis, 1965; Kaplan & Goldsen, 1965; Krippendorff, 1980; Stone, Dunphy, Smith, & Ogilvie, 1966). Krippendorff (1980) highlights the importance of reliability and validity by defining content analysis as "a research technique for making replicable and valid inferences from data to their context" (p. 21). The Joint Committee (1981, pp. 116-123) and the ERS (1982, pp. 13-14) also have standards that specify evaluations should be concerned with both reliability and validity, particularly when delineating and obtaining information. This obviously applies to any evaluation that also uses content analysis methods.


Reliability. Table 4 is used to summarize the classifications of reliability used by the APA/AERA/NCME and the content analysis sources in terms of the type of agreement they best represent. A brief description of each type of reliability follows.

Test-retest agreement represents comparisons over time in general. Stability emphasizes an individual's self-agreement over time.
Table 4
Classifications of Reliability by Type of Agreement Represented

<table>
<thead>
<tr>
<th>Type of agreement</th>
<th>APA/AERA/NCME</th>
<th>Holsti</th>
<th>Krippendorff</th>
<th>Stone et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-retest</td>
<td>Comparisons over time</td>
<td></td>
<td>Stability</td>
<td>Category stability</td>
</tr>
<tr>
<td>Inter-rater</td>
<td>Individual reliability</td>
<td>Reproducibility</td>
<td>Coder reliability</td>
<td></td>
</tr>
<tr>
<td>Test-standard</td>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intra-unit</td>
<td>Unit reliability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater-group</td>
<td>Individual reliability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intra-category</td>
<td>Internal consistency</td>
<td>Category reliability</td>
<td>Single category reliability</td>
<td>Category consistency</td>
</tr>
<tr>
<td>Inference-inference</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form-form</td>
<td>Comparability of forms</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Category stability emphasizes a category's total score agreement over time.

Inter-rater agreement reflects the extent to which the pool of raters agrees on the collection of category ratings. This has been called individual reliability, reproducibility and coder reliability by the content analysis authors.

Test-standard agreement reflects the extent to which the pool of raters agrees with an external standard. Krippendorff calls this accuracy. To the extent the standard is considered to be "correct" or "true" he considers it to be a measure of validity (1980, p. 131).

Intra-unit agreement reflects the extent to which the pool of raters agrees on the coding of a single unit. This has been called unit reliability.

Rater-group agreement reflects the extent to which any single rater agrees with the remainder of the pool of raters. This has been called individual reliability by Krippendorff.

Intra-category agreement reflects the extent to which all the units in a single category represent the same concept. This has been called internal consistency, category reliability, single category reliability, and category consistency.

Intra-category group agreement is the same as above except more than one and fewer than all of the categories are simultaneously considered. This has been called conditional reliability.

Inference-inference agreement reflects the extent to which the pool of raters agrees on the conclusions that can be drawn from the analysis. This has been called interpretive reliability.
Form–form agreement reflects the extent to which administrations of parallel test forms produce comparable results. The APA/AERA/NCME calls this comparability of forms while no content analysis sources include this type of reliability.

Validity. Table 5 is used to summarize the classifications of validity used by the APA/AERA/NCME and the content analysis sources. Two of the sources simply adopt the APA/AERA/NCME classifications. Krippendorff relabels, subdivides, and groups the types of validity in terms of their orientation toward data, product, or process.

Content validity represents the extent to which a coding system and sample of units is representative of the universe of possible units. Krippendorff subdivides the representativeness of the coding system and sample as semantical validity and sampling validity.

Predictive validity represents the extent to which the results of a content analysis can be used to predict an event in the future. Concurrent validity represents the extent to which the results of a content analysis can be substituted for a different analytical procedure. Krippendorff calls both of these predictive validity. This is not consistent with the APA/AERA/NCME interpretation of predictive validity.

Construct validity represents the extent to which theoretical constructs have been substantiated. It requires multiple methods over multiple studies. Krippendorff includes two methods, convergent validation and discriminant validation, under the term correlational validity. Convergent validation represents the extent to which similar variables are highly and positively correlated. Discriminant
Table 5
Classifications of Validity

<table>
<thead>
<tr>
<th>APA/AERA/NCME</th>
<th>Holstl</th>
<th>Krippendorff</th>
<th>Stone et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data oriented</td>
<td>Content</td>
<td>Content</td>
<td>Semantical</td>
</tr>
<tr>
<td>Content</td>
<td>Content</td>
<td>Sampling</td>
<td>Content</td>
</tr>
<tr>
<td>Product oriented</td>
<td>Construct</td>
<td>Construct</td>
<td>Correlational</td>
</tr>
<tr>
<td>Predictive</td>
<td>Predictive</td>
<td>Predictive</td>
<td>Predictive</td>
</tr>
<tr>
<td>Concurrent</td>
<td>Concurrent</td>
<td>Predictive</td>
<td>Concurrent</td>
</tr>
<tr>
<td>Process oriented</td>
<td>Construct</td>
<td>Construct</td>
<td>Construct</td>
</tr>
</tbody>
</table>


Validation represents the extent to which different variables are weakly correlated or negatively correlated. Krippendorff also calls construct validity the extent to which an analytical process parallels or mimics relations in the context where the data were created.

Summary

Standards of quality are important to the fields of evaluation and content analysis. However, evaluation standards are more formal than content analysis standards, even though these standards are best thought of as still emerging from multiple perspectives.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
In addition, evaluation standards are a superset of content analysis standards—they both include expectations of reliability and validity while evaluation standards encompass a much wider range of expectations as well. Because of this, evaluation standards should always be applied whenever a study involves both characterizations and appraisals. If the study involves only descriptions and impartial conclusions, it does not constitute an evaluation. In this case, the standards of reliability and validity alone might suffice.

**Working Definitions of Evaluation and Content Analysis**

The link between evaluation and content analysis can now be established through working definitions that emphasize the similarities and differences between them in terms of information, actions, and standards of quality. Before the definitions themselves are presented, the key concepts on which they are based are reviewed.

First, evaluation and content analysis information focuses on different underlying contrasts. Evaluation information focuses on the contrast between knowledge and value statements. As a result, such information has been called characterizations and appraisals. Content analysis information focuses on the contrast between observations and logic. As a result, such information has been called descriptions and conclusions.

Second, evaluation and content analysis efforts involve the same basic actions. Four of these actions are related to processing some kind of information. They include delineating, obtaining, providing, and applying information. Two of these actions are related to the
total effort. They include managing and evaluating the evaluation or content analysis effort.

Third, high quality is important to both evaluation and content analysis, but the actual standards of quality are highly informal in content analysis and still emerging in evaluation. Because of this, normative definitions that simply draw attention to the issue of quality will be more durable than those that specify particular expectations of quality. Such definitions are sufficient here.

Based on the above considerations, working definitions of evaluation and content analysis with comparable grammatical structures follow. Good evaluation is the high quality process of delineating, obtaining, providing, and applying characterizations and appraisals about some object; and managing and evaluating the evaluation. Good content analysis is the high quality process of delineating, obtaining, providing, and applying descriptions and conclusions about some object; and managing and evaluating the content analysis.

An Overview of Evaluation

Several approaches to conducting evaluation efforts have been developed over the years. Many of these approaches make truly unique contributions to solving important problems, while others are little more than old goods in a new package. Classification systems intended to sort out unique approaches from repackaged goods are presented here to help identify some basic schools of thought for conducting an evaluation. After these approaches are identified, they are summarized in terms of a few important attributes.
Over the past twenty years or so, the number of alternative approaches to conducting evaluation efforts has skyrocketed. Factors such as the Elementary and Secondary Education Act (ESEA) of 1965 that required educators to evaluate their efforts and results, and the growing public concern for accountability of human service programs contributed to this growth (Borich & Jemelka, 1982, p. 4). Many sources that provide historical accounts of the proliferation of evaluation approaches are available. Works emphasizing educational evaluation include Baker (1980); Cronbach et al. (1980); Guba and Lincoln (1981); and Madaus, Stufflebeam, and Scriven (1983). Works on human service evaluation include Attkisson and Broskowski (1978), and Flaherty and Morell (1978). Borich and Jemelka (1982) provide a history of evaluation covering education and human services. Systems for classifying evaluation approaches are discussed next.

Classification of Approaches

A number of authors have provided classifications of various types of evaluation approaches (e.g., Borich & Jemelka, 1982, pp. 7-18; Guba & Lincoln, 1981, pp. 1-38; House, 1978; Patton, 1981, pp. 186-193; Popham, 1975, pp. 20-44; Stake, 1974; Stufflebeam & Webster, 1980; Worthen & Sanders, 1973, pp. 209-217). The number of general approaches presented in these sources covers a rather large range. For example, Guba and Lincoln (1981, p. 38) contend all evaluation approaches can ultimately be subsumed under one approach—responsive evaluation. At the other extreme, Patton (1981, pp. 186-193)—with tongue in cheek—presents a "beginning list" of 132 approaches. Most
other authors present around ten generalized evaluation approaches.

Two classifications of evaluation approaches (House, 1978; Stufflebeam & Webster, 1980) are of particular interest for this study because they include a manageable number of approaches, and they group these approaches in terms of underlying principles similar to those used in the preceding section of this chapter. The general structures of these classification systems are discussed first. The structures are then combined to present a more refined classification of fifteen evaluation approaches. A summary of these approaches is presented in the next section.

House (1978) considers all major evaluation approaches to be based on a common ideology, liberal democracy. Important principles of this ideology include freedom of choice, the uniqueness of the individual, and empirical inquiry. He also contends they are all based on subjectivist ethics, in which ethical conduct is based on the subjective or intuitive experience of an individual or group. One form of subjectivist ethics is utilitarian, in which "the good" is determined by what maximizes some single, explicit interpretation of happiness for society as a whole. Another form of subjectivist ethics is intuitionist/pluralist, in which no single interpretation of "the good" is assumed and these interpretations need not be explicitly stated nor justified.

These ethical positions have corresponding epistemologies—philosophies of obtaining knowledge. The objectivist epistemology is associated with the utilitarian ethic. In general, it is used to acquire knowledge capable of external verification (intersubjective
agreement) through publicly inspectable methods and data. The subjectivist epistemology is associated with the intuitionist/pluralist ethic. It is used to acquire new knowledge based on existing personal knowledge and experiences that are (explicit) or are not (tacit) available for public inspection.

House further divides each epistemological approach by two main political perspectives. Approaches can take an elite perspective, focusing on the interests of managers and professionals. They can also take a mass perspective, focusing on consumers and participatory approaches.

Stufflebeam and Webster (1980) place approaches into one of three groups according to their orientation toward the role of values, an ethical consideration. The political orientation promotes a positive or negative view of an object regardless of what its value might actually be. They call this pseudo-evaluation. The questions orientation includes approaches that might or might not provide answers specifically related to the value of an object. They call this quasi-evaluation. The values orientation includes approaches primarily intended to determine the value of some object. They call this true evaluation.

Table 6 is used to classify fifteen evaluation approaches in terms of epistemology, major perspective (from House), and orientation (from Stufflebeam & Webster). When considered simultaneously, these three dimensions produce twelve cells. Only seven of the cells contain approaches, although all four true evaluation cells contain at least one approach.
<table>
<thead>
<tr>
<th>Epistemology</th>
<th>Major perspective</th>
<th>Orientation</th>
<th>Questions (Quasi-evaluation)</th>
<th>Values (True evaluation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectivist</td>
<td>Elite (Managerial)</td>
<td>Politically controlled Public relations</td>
<td>Experimental research Management information systems Testing programs Objectives-based Content analysis</td>
<td>Decision-oriented Policy studies</td>
</tr>
<tr>
<td>Objectivist</td>
<td>Mass (Consumers)</td>
<td>Accountability</td>
<td>Consumer-oriented</td>
<td></td>
</tr>
<tr>
<td>Subjectivist</td>
<td>Elite (Professional)</td>
<td></td>
<td>Accreditation/ Certification Connoisseur</td>
<td></td>
</tr>
<tr>
<td>Subjectivist</td>
<td>Mass (Participatory)</td>
<td></td>
<td>Adversary Client-centered</td>
<td></td>
</tr>
</tbody>
</table>

Two pseudo-evaluation approaches, politically controlled and public relations studies, are represented. They are based on an objectivist epistemology from an elite perspective.

Six quasi-evaluation approaches use an objectivist epistemology. Five of them—experimental research, management information systems, testing programs, objectives-based studies, and content analysis—take an elite perspective. Accountability takes a mass perspective.

Seven true evaluation approaches are included. Two approaches, decision-oriented and policy studies, are based on an objectivist epistemology from an elite perspective. Consumer-oriented studies are based on an objectivist epistemology from a mass perspective. Two approaches—accreditation/certification and connoisseur studies—are based on a subjectivist epistemology from an elite perspective. Finally, adversary and client-centered studies are based on a subjectivist epistemology from a mass perspective. A summary of the fifteen approaches is presented next.

Summary of Approaches

The preceding section was used to distinguish between fifteen evaluation approaches in terms of their epistemology, major perspective, and orientation to values. This classification resulted in a twelve-celled matrix of which seven cells contain at least one entry. Five of the cells contain more than one entry. This section is intended to summarize each of the fifteen approaches in enough detail so that those placed in the same cell of Table 6 can be distinguished from each other.
Table 7 is used to summarize each approach in terms of four attributes—organizer, purpose, strengths, and weaknesses. The organizer represents the main considerations or cues practitioners use to organize a study. The purpose represents the desired outcome for a study at a very general level. Strengths and weaknesses represent other attributes that should be considered when deciding whether to use the approach for a particular study. Space considerations preclude most of the information in this table from being duplicated in the text. Instead, the following narrative highlights differences between approaches that are grouped into the same cell of Table 6. Sources for further reading on each approach are also presented here.

**Pseudo-Evaluation.** Only one of four pseudo-evaluation cells contains any entries. Politically controlled and public relations studies are based on an objectivist epistemology from an elite perspective. Although both of these approaches seek to misrepresent value interpretations about some object, they go about it a bit differently. Information obtained through politically controlled studies is released or withheld to meet the special interests of the holder. Public relations information is used to paint a positive image of an object regardless of the actual situation. Neither of these approaches is acceptable evaluation practice, although the seasoned reader can surely think of a few examples where they have been used.

**Objectivist, Elite, Quasi-Evaluation.** Five approaches are contained in this cell. As a group, they represent a highly respected collection of disciplined inquiry approaches. They are considered
<table>
<thead>
<tr>
<th>Approach</th>
<th>Organizer</th>
<th>Purpose</th>
<th>Key strengths</th>
<th>Key weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politically controlled</td>
<td>Threats</td>
<td>Get, keep or increase influence, power, or money.</td>
<td>Secures evidence advantageous to the client in a conflict.</td>
<td>Violates the principle of full &amp; frank disclosure.</td>
</tr>
<tr>
<td>Public relations</td>
<td>Propaganda needs</td>
<td>Create positive public image.</td>
<td>Secures evidence most likely to bolster public support.</td>
<td>Violates the principles of balanced reporting, justified conclusions, &amp; objectivity.</td>
</tr>
<tr>
<td>Experimental research</td>
<td>Causal relationships</td>
<td>Determine causal relationships between variables.</td>
<td>Strongest paradigm for determining causal relationships.</td>
<td>Requires controlled setting, limits range of evidence, focuses primarily on results.</td>
</tr>
<tr>
<td>Management evidence systems</td>
<td>Scientific efficiency</td>
<td>Continuously supply evidence needed to fund, direct, &amp; control programs.</td>
<td>Gives managers detailed evidence about complex programs.</td>
<td>Human service variables are rarely amenable to the narrow, quantitative definitions needed.</td>
</tr>
<tr>
<td>Approach</td>
<td>Organizer</td>
<td>Purpose</td>
<td>Key strengths</td>
<td>Key weaknesses</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------------------------</td>
<td>----------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Testing programs</td>
<td>Individual differences</td>
<td>Compare test scores of individuals &amp; groups to selected norms.</td>
<td>Produces valid &amp; reliable evidence in many performance areas. Very familiar to public.</td>
<td>Data usually only on testee performance, overemphasizes test-taking skills, can be poor sample of what is taught or expected.</td>
</tr>
<tr>
<td>Objectives-based</td>
<td>Objectives</td>
<td>Relate outcomes to objectives.</td>
<td>Common sense appeal, widely uses, used behavioral objectives &amp; testing technologies.</td>
<td>Leads to terminal evidence often too narrow to provide basis for judging the value of a program.</td>
</tr>
<tr>
<td>Content analysis</td>
<td>Content of a communication</td>
<td>Describe &amp; draw conclusions about a communication.</td>
<td>Allows for unobtrusive analysis of large volumes of unstructured, symbolic materials.</td>
<td>Samples may be unrepresentative yet overwhelming in volume. Analysis design often overly simplistic for question.</td>
</tr>
<tr>
<td>Approach</td>
<td>Organizer</td>
<td>Purpose</td>
<td>Key strengths</td>
<td>Key weaknesses</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------</td>
<td>-------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Decision-oriented</td>
<td>Decisions</td>
<td>Provide a knowledge &amp; value base for making &amp; defending decisions.</td>
<td>Encourages use of evaluation to plan &amp; implement needed programs. Helps justify decisions about plans &amp; actions.</td>
<td>Necessary collaboration between evaluator &amp; decision-maker provides opportunity to bias results.</td>
</tr>
<tr>
<td>Policy studies</td>
<td>Broad issues</td>
<td>Identify and assess potential costs &amp; benefits of competing policies.</td>
<td>Provides general direction for broadly focused actions.</td>
<td>Often corrupted or subverted by politically motivated actions of participants.</td>
</tr>
<tr>
<td>Consumer-oriented</td>
<td>Generalized needs &amp; values, effects</td>
<td>Judge the relative merits of alternative goods &amp; services.</td>
<td>Independent appraisal to protect practitioners &amp; consumers from shoddy products &amp; services.</td>
<td>Might not help practitioners do a better job. Requires credible &amp; competent evaluator.</td>
</tr>
<tr>
<td>Accreditation/certification</td>
<td>Standards &amp; guidelines</td>
<td>Determine if institutions, programs, &amp; personnel should be approved to perform specified functions.</td>
<td>Helps public make informed decisions about quality of organizations &amp; qualifications of personnel.</td>
<td>Standards &amp; guidelines typically emphasize intrinsic criteria to the exclusion of outcome measures.</td>
</tr>
<tr>
<td>Approach</td>
<td>Organizer</td>
<td>Purpose</td>
<td>Key strengths</td>
<td>Key weaknesses</td>
</tr>
<tr>
<td>------------------</td>
<td>-------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Connoisseur</td>
<td>Critical guideposts</td>
<td>Critically describe, appraise, &amp; illuminate an object.</td>
<td>Exploits highly developed expertise on subject of interest. Can inspire others to more insightful efforts.</td>
<td>Dependent on small number of experts, making evaluation susceptible to subjectivity, bias, and corruption.</td>
</tr>
<tr>
<td>Adversary</td>
<td>&quot;Hot&quot; issues</td>
<td>Present the pros &amp; cons of an issue.</td>
<td>Ensures balanced presentations of represented perspectives.</td>
<td>Can discourage cooperation, heighten animosities.</td>
</tr>
<tr>
<td>Client-centered</td>
<td>Specific concerns &amp; issues</td>
<td>Foster understanding of activities &amp; how they are valued in a given setting &amp; from a variety of perspectives.</td>
<td>Practitioners are helped to conduct their own evaluation.</td>
<td>Low external credibility, susceptible to bias in favor of participants.</td>
</tr>
</tbody>
</table>

Note. Adapted and condensed primarily from House (1978) and Stufflebeam & Webster (1980).
quasi-evaluation approaches because particular studies can legitimately focus only on questions of knowledge without addressing any questions of value. Such studies are, by definition, not evaluations. Using the terminology developed in the first section of this chapter, these approaches can produce characterizations without producing appraisals, although specific studies can produce both. Each of these approaches serves its intended purpose well. They are discussed roughly in order of the extent to which they approach the objectivist ideal.

Experimental research is the best approach for determining causal relationships between variables. The potential problem with using this as an evaluation approach is that its highly controlled and stylized methodology may not be sufficiently responsive to the dynamically changing needs of most human service programs. Important contributors to the experimental research approach include Lindquist (1953), and Cook and Campbell (1979).

Management information systems (MIS's) can give detailed information about the dynamic operations of complex programs. However, this information is restricted to readily quantifiable data usually available at regular intervals. Contributors to the MIS approach include Cook (1966), Kauffman (1969), and Rivlin (1971).

Testing programs are familiar to just about anyone who has attended school, served in the military, or worked for a large company. These programs are good at comparing individuals or groups to selected norms in a number of subject areas or to a set of standards of performance. However, they only focus on testee performance and
they might not adequately sample what is taught or expected. CONTRIBUTORS to the testing program approach include Lindquist (1951), Ebel (1965), Hambleton and Swaminathan (1985), and Thorndike (1971).

Objectives-based approaches relate outcomes to prespecified objectives, allowing judgments to be made about their level of attainment. Unfortunately, the objectives are often not proven to be important or they focus on outcomes too narrow to provide the basis for determining the value of an object. CONTRIBUTORS to this approach include Tyler (1949); Bloom, Englehart, Furst, Hill, and Krathwohl (1956); Hammond (1973); Kiresuk and Lund (1978); Krathwohl, Bloom, and Masia (1964); Metfessel and Michael (1967); Popham (1969); and Provus (1971).

Content analysis was not included in the original classifications of evaluation approaches used for this section (House, 1978; Stufflebeam & Webster, 1980). It was added to the list to place it in the context of commonly accepted evaluation approaches. It is a quasi-evaluation approach because content analysis judgments need not be based on value statements. Instead, they can be based on knowledge. Such content analyses are not evaluations. On the other hand, when content analysis judgments are based on values, such studies are evaluations. This approach is discussed in detail in a later section. Key contributors to content analysis include Berelson (1952); Krippendorff (1980); Lasswell, Leites, and Associates (1965); Holsti (1969); and Stone et al. (1966).

Objectivist, Mass, Quasi-Evaluation. Accountability is the only approach assigned to this cell. It is popular with constituents
because it is intended to provide an accurate accounting of results that can improve the quality of products and services. However, this approach can quickly turn practitioners and consumers into adversaries when implemented in a heavy-handed fashion. The leading contributor to the accountability approach is Lessinger (1970).

**Objectivist, Elite, True Evaluation.** The evaluation approaches in this cell are of two basic types. They include decision-oriented studies and policy studies.

Decision-oriented studies are designed to provide a knowledge base for making and defending decisions. This approach usually requires the close collaboration between an evaluator and decision-maker, allowing it to be susceptible to corruption and bias. Contributors to this approach include Cronbach (1963), Stufflebeam et al. (1971), and Alkin (1969).

Policy studies provide general guidance and direction on broad issues by identifying and assessing potential costs and benefits of competing policies. The drawback is these studies can be corrupted or subverted by the politically motivated actions of the participants. Contributors to the policy study approach include Coleman et al. (1966), Jenks et al. (1972), and Clark (1965).

**Objectivist, Mass, True Evaluation.** One approach has been placed in this cell, consumer-oriented studies. This approach is used to judge the relative merits of goods and services based on generalized needs and values, along with a comprehensive range of effects. However, this approach does not necessarily help practitioners improve their work, and it requires a very good and credible
evaluator to do it well. The single most important contributor to
the description and practice of this evaluation approach is Scriven
(1967, 1974b).

Subjectivist, Elite, True Evaluation. This cell contains three
approaches, accreditation/certification, policy studies, and connois-
seur studies. They represent one of two groups of approaches that
use a subjectivist epistemology.

Accreditation/certification programs are based on self-study and
peer review of organizations, programs, and personnel. They draw on
the insights, experience, and expertise of qualified individuals who
use established guidelines to determine if the applicant should be
approved to perform specified functions. However, attributes of
applicants and the processes they perform are often overemphasized in
relation to measures of outcomes or effects. Examples of accredita-
tion boards include the Commission on Accreditation of Rehabilitation
Facilities, the Joint Commission on Accreditation of Hospitals, and
the North Central Association of Secondary Schools and Colleges.

Connoisseur studies use the highly refined skills of individuals
intimately familiar with the subject of the evaluation to critically
characterize and appraise it. This approach can help others see
programs in a new light, but it is difficult to find a qualified and
unbiased connoisseur. Contributors to this approach include Eisner
(1975), Guba (1978), and Sanders and Hershiser (1976).

Subjectivist, Mass, True Evaluation. This final cell contains
two approaches, adversary and client-centered studies. They use a
subjectivist epistemology from multiple audience perspectives.
The adversary approach focuses on drawing out the pros and cons of controversial issues through quasi-legal proceedings. This helps ensure a balanced presentation of different perspectives on the issues, but it is also likely to discourage later cooperation and heighten animosities between contesting parties if "winners" and "losers" emerge. Contributors to this approach include Owens (1971) and Wolf (1973).

The last approach covered includes client-centered studies. They address specific concerns and issues of practitioners and other clients of the study in a particular setting. These studies help people understand the activities and values involved from a variety of perspectives. However, this responsive approach can lead to low external credibility and a favorable bias toward those who participated in the study. Contributors to this approach include Stake (1967), Guba (1978), Guba and Lincoln (1981), and Rippey (1973).

A General Model for Conducting an Evaluation Effort

The first area of discussion in this chapter focused on the conceptual relationships between evaluation and content analysis. The second area of discussion focused on alternative approaches to conducting evaluations and identified content analysis as a quasi-evaluation approach. This third area of discussion focuses on a general model for conducting an evaluation effort that is consistent with the previous discussions; accommodates content analysis experts' views on the uses, tasks, and computer implementations of content analysis; and provides the framework for an experimental study on how
microcomputers can be used to improve content analyses of responses to open-ended survey questions used in evaluation efforts.

The general model for conducting an evaluation effort reflects the key relationships between the information, action, and standards of quality components discussed earlier. It can also be thought of as a graphic version of the working definition of evaluation. The model is presented in Figure 2.

Figure 2. The General Model for Conducting an Evaluation Effort
In this context, evaluation is best thought of as a process that includes six actions. Four of them—delineating, obtaining, providing, and applying—focus on information. This information includes characterizations and appraisals about some object.

The other two basic actions—managing and evaluating—focus on the overall effort. Managing the effort begins at some time during the early delineating activities and it ends at some phase of applying the information. Exactly when management of the effort begins and ends is dependent on the particular evaluation approach used. Evaluating the effort, or meta-evaluation, can be used to scrutinize events that occurred well before and after the official time period of the study, although most of the focus is usually placed on the official information processing actions and their consequences. It should also be noted the working definition stipulates all the actions and information should be of high quality, including managing and evaluating the effort.

While Figure 2 is used to show the basic relationships between the information, actions, and standards of quality for an evaluation effort, Figure 3 better illustrates its dynamic nature. This figure is used to focus on the decision network for delineating, obtaining, providing, and applying evaluative information—characterizations and appraisals. The network itself is represented in the lower portion of the figure, while the upper portion is used to remind the reader these actions still need to be managed and evaluated.

From a logical perspective, any path that follows the flow of the arrows is possible, although certain patterns are more likely
Figure 3. Decision Network Using the General Model for Conducting an Evaluation Effort

than others. For example, if one were to ask if an evaluation should be conducted at all but immediately answered no, the lower path that bypasses the heart of the process is followed.

This network can also accommodate two ideal types of evaluation information processing patterns that are conceptually incompatible but are probably never found in their "pure" forms. These patterns of information processing are often called "preordinate" and "responsive" in the evaluation literature (e.g., Guba & Lincoln, 1981; Stake, 1975). The preordinate pattern is exemplified by the experimental research approach presented earlier. In this pattern, all delineating activities are completed before any of the obtaining activities begin. In a like manner, all obtaining activities are completed before providing information begins, which is completed
before applying information begins. In the responsive pattern, exemplified by the client-centered approach discussed earlier, several iterations of delineating, obtaining, providing, and applying information about each aspect and subplot of an evaluation are undertaken before the effort as a whole is completed. In reality, however, even preordinate efforts often need to follow side issues or return to a previous stage of a study to modify work already completed; and responsive efforts often complete substantial portions of a particular type of action before moving on to the next stage of the study.

In addition, this network is hierarchically recursive in nature. In other words, in order to complete a major information processing action, supporting actions often need to be completed first. For example, before providing an appraisal of a school district's accountability system to the school board, it is first necessary to characterize how various interest groups view the system.

In summary, the general model for conducting an evaluation effort can be used to show the logical relationships between its information, action, and standards of quality components. It can also accommodate many different patterns of processing characterizations and appraisals of some object.

Because content analysis is a quasi-evaluation approach, the general model also applies to such efforts. The remainder of this chapter focuses on that approach in terms of its general uses, tasks, and the performance of certain of these tasks with the help of microcomputers.
An Overview of Content Analysis

Around the turn of the century, the use of content analysis as a serious approach to studying the world in general and the nature of communication in particular began to increase. Because most early studies were concerned with describing daily newspapers in terms of the amount of coverage they gave to different subject areas (e.g., Speed, 1893), the approach was then called "quantitative newspaper analysis" (Krippendorff, 1980, p. 14). However, the popularity of the approach soon began to grow into different fields (e.g., military intelligence, psychotherapy, history, anthropology, education, literature, & linguistics) and different media (e.g., books, documents, correspondence, movies, radio, TV, & photography). Historical accounts of this growth in the field are available from a number of sources (e.g., Barcus, 1959; Berelson, 1952, pp. 21-26; Carney, 1972, pp. 26-36; Holsti, 1969, pp. 20-23; Krippendorff, 1980, pp. 13-20; Stone et al., 1966, pp. 21-44). This discussion is organized into three basic topics: (1) a summary of the uses of content analysis, (2) a summary of the tasks of content analysis, and (3) a summary of the ways computers in general and microcomputers in particular can be used to help perform some of these tasks.

Classification of Uses

Several authors have provided classifications of content analysis and its uses. One of the earliest classification systems was developed by Janis (1943). This classification system also appears
in a book devoted to using content analysis methods to help study the field of politics (Janis, 1965). The three main types of content analysis are described as:

1. **Pragmatical Content Analysis**—procedures which classify signs according to their probable causes or effects (e.g., counting the number of times that something is said which is likely to have the effect of producing favorable attitudes toward Germany in a given audience).

2. **Semantical Content Analysis**—procedures which classify signs according to their meaning (e.g., counting the number of times Germany is referred to, irrespective of the particular words that may be used to make the reference).
   - (a) **designations analysis**—provides the frequency with which certain objects (persons, things, groups or concepts) are referred to, i.e., roughly speaking, subject-matter analysis (e.g., references to German foreign policy).
   - (b) **attribution analysis**—provides the frequency with which certain characterizations are referred to (e.g., references to dishonesty).
   - (c) **assertions analysis**—provides the frequency with which certain objects are characterized in a particular way, i.e., roughly speaking, thematic analysis (e.g., references to German foreign policy as dishonest).

3. **Sign-vehicle analysis**—procedures which classify content according to the psychophysical properties of the signs (e.g., counting the number of times the word "Germany" appears). (p. 57)

Berelson (1952) places seventeen uses of content analysis into four basic groups, one of which has two subgroups. They are: "[1] characteristics of content . . . , [a] substance . . . , [b] form . . . ; [2] producers of content . . . ; [3] audiences of content . . . ; and [4] effects of content" (pp. 27-29).

Holsti (1969) draws from the above sources to group fifteen uses into three basic categories. His three categories are to: "[1] describe characteristics of communication . . . , [2] make inferences . . . . . ."
as to the antecedents of communication . . . , and [3] make inferences as to the effects of communication" (p. 26). He also places these uses into the context of the communication paradigm and states:

Content analysis is always performed on the message, be it a novel, diplomatic note, editorial, diary, or speech. The results of content analysis may, however, be used to make inferences about all other elements of the communication process. To the classical formulation of these questions—"who says what, to whom, how, and with what effect?" (Lasswell, Lerner, and Pool, 1952, p. 12)—we shall add one more: "why?" (p. 24)


Because Holsti's classification system accommodates almost all of the specific uses presented by the other authors, its underlying structure—the communication paradigm—is familiar to a general audience, and it is compatible with the working definition of content analysis; it is used as the main organizer for the discussion on the uses of content analysis. The contributions of the other authors that expand this basic framework are added where appropriate.

**Summary of Uses**

Table 8 is used to summarize fifteen uses of content analysis in terms of their general purpose, element of the communication paradigm.
Table 8

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Element</th>
<th>Question</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Make inferences about the antecedents of communications</td>
<td>Source</td>
<td>Who?</td>
<td>Answer questions of disputed authorship</td>
</tr>
<tr>
<td></td>
<td>Encoding process</td>
<td>Why?</td>
<td>Secure political &amp; military intelligence</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Analyze traits of individuals</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Infer cultural aspects &amp; change</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Provide legal &amp; evaluative evidence</td>
</tr>
<tr>
<td>Describe &amp; make inferences about the characteristics of communications</td>
<td>Channel</td>
<td>How?</td>
<td>Analyze techniques of persuasion</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Analyze style</td>
</tr>
<tr>
<td></td>
<td>Message</td>
<td>What?</td>
<td>Describe trends in communication content</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Relate known characteristics of sources to messages they produce</td>
</tr>
<tr>
<td></td>
<td>Recipient</td>
<td>To whom?</td>
<td>Describe patterns of communication</td>
</tr>
<tr>
<td>Make inferences about the consequences of communications</td>
<td>Decoding process</td>
<td>With what effect?</td>
<td>Measure readability</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Analyze the flow of information</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Assess responses to communications</td>
</tr>
</tbody>
</table>

to which they apply, and the general question they are intended to answer. In Figure 4, this same information is graphically presented in relation to the working definition of content analysis. Thus, using action forms of the definition’s information components, content analysis is a process for describing messages and making inferences about any element of the communication process. Managing and evaluating the effort are also necessary if high quality results are desired. The uses of content analysis that apply to the antecedents, characteristics, and consequences of communications are discussed next. However, due to space considerations, references to specific content analysis studies and methods are kept to a minimum. Readers interested in learning more about particular uses in specific contexts are referred to Berelson (1952) and Holsti (1969).

**Antecedents of Communications**

Making inferences about the antecedents of communications is a problem in pragmatics—identifying the relationships of signs (messages) to those who produced them. Because of differences in the ways people express their feelings, intentions, and other attributes, inferences about the antecedent causes of messages drawn solely from content data cannot be self-validating. Thus, however precise our measures of communication content, it is hazardous indeed to assume, without corroborating evidence from independent, noncontent data, that inferences about the author may be drawn directly from content data. (Holsti, 1969, p. 32)

Because of this problem, two basic types of comparisons—direct and indirect—are used to draw inferences about the antecedents of
Figure 4. Uses of Content Analysis in Terms of the Working Definition and the Communication Paradigm
messages. With direct comparisons, a content variable from a given source (e.g., aggressive threats of a specific terrorist group) is compared to a behavioral variable of the same type for the same source (e.g., actual terrorist attacks by that group).

Indirect comparisons, on the other hand, relate a content variable from one source to the same content variable from a different source. The different source can be a completely different individual or the same individual at a different point in time. In addition, the content variable from the other source must have been directly compared to a related behavioral variable. For example, comparing the coverage of one textbook to that of another considered to be culturally biased can be used to make inferences about the biases of the first textbook. An example of an indirect comparison for a single individual is when a mother sends her four-year-old to bed at 7:30 P.M. after hearing a series of complaints—knowing full well things will only get worse if the child is allowed to stay up any later.

Any inferences about the antecedents of a communication must also be based on either a representational or instrumental model of communication. In the representational model,

the important point about the communication is what is revealed by the content of the [words] present in it; that is, something in the words of the message may have indicatorial validity regardless of the circumstances, and it is at the message that the analyst looks. (Pool, 1959, p. 3)

The instrumental model "signifies that the important point is not what the message says on the face of it but what it conveys, given
its context and circumstances" (p. 3). For a further discussion of the representational and instrumental models, see Pool (1959, chaps. 1-3, 7) and Mitchell (1967).

Two questions—"who?" and "why?"—apply to making inferences about the antecedents of communications. "Who?" refers to the source of a message. Answering questions of disputed authorship is the only use of content analysis included here. "Why?" refers to the encoding process for a message. "What are the meanings, associations, values, motives, or intentions of the communicator that can be inferred from his [or her] messages?" (Holsti, 1969, p. 32). The uses of content analysis in this group are to: (a) secure potential military intelligence, (b) analyze psychological traits of individuals, (c) infer cultural aspects and change, and (d) provide legal or evaluative evidence.

Two uses, (b) and (d), are probably more interesting to evaluators than the others. For example, surveys can be used to collect evidence about attitudes different interest groups have concerning the object of an evaluation. When they use open-ended questions—questions for which people are asked to respond in their own words—content analysis methods can be used to help summarize the narrative responses (Berelson, 1952, pp. 53-56; Caulley, 1983, p. 22). Methods for analyzing such responses and some suggestions for when open-ended questions should be used at all are discussed next.

Evaluative assertion analysis (Osgood, 1959; Osgood, Saporta, & Nunnally, 1956) can be used to determine the attitudes of different groups about a particular evaluation object, such as a public
school's accountability system. "The purpose of evaluative assertion analysis is to extract from messages the evaluations being made [by the source] of significant concepts" (Osgood, 1959, p. 42). The crucial first phase of evaluative assertion analysis involves transforming the original message into a series of evaluative assertions ("neutral" assertions are not included in this particular type of analysis) with one of two possible grammatical structures: (1) attitude object / verbal connector / common meaning term, or (2) attitude object_1 / verbal connector / attitude object_2. For example, the sentence, "All teachers despise the dogmatic accountability system," is translated to read: (1) All teachers / despise / the accountability system, and (2) The accountability system / is / dogmatic. Quantitative techniques are then performed on these transformed assertions. The final result is a numerical indicator of the direction and intensity of all evaluative assertions about each attitude object. However, the transformed assertions themselves constitute a type of "qualitative" information that can also be very useful. Unfortunately, because of the labor-intensive nature of the method, Osgood (1959) concludes, "This method is more likely to find use as a research tool than in practically oriented areas" (p. 53). Evaluation qualifies as one of these practically oriented areas.

Another method developed by Osgood for summarizing the attitudes of reference groups is the semantic differential technique (Osgood, 1962; Osgood Suci, & Tannenbaum, 1957; Snider & Osgood, 1969). This method is not used to analyze narrative data. Instead, concepts or attitude objects are related in terms of a number of bipolar scales.
like "aloof-responsive," or "fair-biased." These attitude objects are then represented in three empirically derived dimensions of meaning—evaluation (e.g., good-bad or positive-negative), potency (e.g., strong-weak or hard-soft), and activity (e.g., active-passive or slow-fast)—(Osgood et al., 1957, pp. 62-63). However, comparing narrative data to these dimensions, particularly evaluation, is possible. For example, responses to open-ended questions can first be coded by the main attitude object they contain and then each response can be scored on a rating scale (e.g., Stanley & Hopkins, 1972, p. 290) that represents the evaluation dimension of the semantic differential (e.g., positive-negative). Depending on the precision desired, the scale can have anywhere from two to seven points. Using a three-point scale (negative-neutral-positive) and the sentence, "All teachers despise the dogmatic accountability system;" the main attitude object is, "the accountability system," and the rating on the evaluation scale is, "negative."

Survey researchers usually distinguish between two basic types of questions—open-ended and forced-choice. Open-ended questions require people to respond in their own words while forced-choice questions require people to choose from a set of predefined alternatives. Obviously, it is the open-ended type of question whose responses must be subjected to content analysis.

Payne (1951) identifies several uses of open-ended questions. Common uses are to: (a) introduce a topic to the respondent, (b) provide background information for interpreting responses to other questions, (c) obtain elaborations of previous responses, (d) elicit
reasons for previous responses, (e) elicit presumed arguments for different sides of an issue, (f) explore knowledge and memory, (g) identify sources of information, (h) obtain factual information, (i) provide preliminary information for drafting forced-choice questions, and (j) provide a source of quotations for final reports (pp. 34-50).

Open-ended and forced-choice questions both have characteristic strengths and weaknesses (e.g., Demaline & Quinn, 1979, pp. 30-31; Downs, Smeyak, & Martin, 1980, pp. 44-48; Payne, 1951, pp. 49-54; Warwick & Linnger, 1975, pp. 132-140). The summary by Demaline and Quinn (1979) is both concise and representative of the other authors. They present the following advantages and disadvantages of open-ended and forced-choice questions:

**Forced-Choice Questions**

**Advantages**
1. It is easier for respondent to answer.
2. Focuses respondent's answer on issues and data of importance to you. Respondents categorize themselves instead of you categorizing them.
3. More questions can be asked because time is saved by the respondent simply checking.
4. Precoded answers are easily analyzed.

**Disadvantages**
1. It requires advance information about possible response categories that may be given.
2. It may bias responses by suggesting answers.
3. It does not allow for diversity and richness in individual expression.

**Open-Ended Questions**

**Advantages**
1. It can easily be formulated without knowing the full range of answers that may be given.
2. It can accommodate questions for which a wide range of different answers will be given.
3. It does not condition or bias the answer as much as the forced-choice question.
Disadvantages
1. It requires the respondent to write a lot. Communication skills may influence the answer, in addition to the other respondent characteristics.
2. The respondent may address different facets of the question in which you may not be interested or may not give complete information in answering the question.
3. Fewer questions can be asked in a questionnaire because answers may be lengthy and time consuming to give.
4. Responses are difficult to analyze. The investigator must devise a coding scheme and then categorize responses based on this scheme. The diversity and richness of responses are usually reduced by this process, and it is time consuming. (pp. 30-31)

Because the advantages and disadvantages of open-ended and forced-choice questions tend to complement each other, suggestions for when either type of question should be used can be given. For example, use forced-choice questions if any of these four conditions apply: (1) it is important the effort or verbal skills of the respondents be kept relatively low, or (2) a clear understanding exists of what the likely or important responses will be, or (3) a large number of questions need to be asked in relation to the time available, or (4) it is important the responses be easy to code for analysis. Use open-ended questions and content analysis if either of these two conditions apply: (1) the full range of likely and valid responses is not known or a wide range of responses is expected, or (2) concern for biasing respondents exists if a set of possible answers is given, and this third condition applies: (3) the skills and time needed for coding potentially complex responses are available. Thus, open-ended questions can be extremely useful, but the effort and content analysis skills needed to properly analyze the responses to them require their judicious use. Payne (1951)
concludes, "Its virtues and its faults all stem from this open feature. Its results are as full of variety as a country store, and just as hard to divide into departments" (p. 54).

Characteristics of Communications

Describing and making inferences about the characteristics of communications are problems in semantics and syntactics, what Jánis (1965) calls semantical content analysis and sign-vehicle analysis (p. 57). Exactly when "describing" becomes "making inferences" is highly dependent on the analyst's position on the objectivist vs. subjectivist issue as discussed under the classification of alternative evaluation approaches. The representational vs. instrumental communication model controversy is a specific example of this general issue. Analysts using the representational model are more likely to "describe" the characteristics of communications, while analysts using the instrumental model are more likely to "make inferences" about the characteristics of communication.

Three basic questions--"how?", "what?", and "to whom?"--apply to describing and making inferences about the characteristics of communications. "How?" refers to the channel of a message. The uses of content analysis in this group are to: (a) analyze techniques of persuasion, and (b) analyze style. "What?" refers to the message itself. The uses grouped under this question are to: (a) describe trends in communication content, (b) relate known characteristics of sources to messages they produce, and (c) compare communication content to standards. The use of content analysis to compare
messages to standards is discussed next. "To whom?" refers to the recipient of the message. The content analysis uses here are to: (a) relate known characteristics of audiences to messages produced for them, and (b) describe patterns of communication.

Berelson (1952), Holsti (1969), and Krippendorff (1980) all partition the uses of content analysis related to standards differently. However, it should be noted they all use the term to mean what is compared to a message, not what is used to judge the quality of the content analysis descriptions, inferences, or processes used to develop them.

Berelson (1952) uses two main categories. The first one, to audit communications against objectives (pp. 43-45), checks a message "against the communicator's own professed objectives" (p. 44). The second category, to construct and apply communication standards (pp. 46-52), compares messages "with the standards of the analyst" (p. 44). He also lists three methods of evaluating communications:

(1) evaluation of performance against such a priori standards as "balance" or "social purpose"; (2) evaluation of performance by comparing one body of content with another (the internal criteria); and (3) evaluation of performance by comparing content with a non-content source (the external criteria). (p. 46)

Holsti (1969) drops the distinction between the source of the standards but maintains the distinction between the three types of standards—a priori, content, and non-content (pp. 53-59). His examples of a priori standards used in actual studies include social norms, responsibility in mass communication, and bias (pp. 54-56).
Content standards are based on "general norms for classes of communi-
cators" (p. 56). Non-content standards include such things as veri-
fied news events to assess newspaper story coverage (p. 54); census
data to assess the distribution of ethnic group members in popular
fiction (p. 58); and expert opinion as the basis to judge depictions
of mental health issues in the general press, or adherence to the
"American creed" in ethnic newspapers (p. 58).

Krippendorf (1980) identifies three types of content analysis
studies that use standards—evaluations, identifications, and audits.

While evaluations assess the degree [emphasis added] to
which something conforms or deviates from a standard,
identifications have a more either/or quality. . . . Au-
dits, too, involve judgments on data relative to a standard
with the additional provision that the standard is pre-
scribed or legitimated by an institution. (p. 39)

Krippendorf's view of identifications most closely resembles
Janis" (1965) categories of semantical and sign-vehicle content
analysis (p. 57). Identifications also reflect the importance of the
role of concept learning in content analysis. "Learning a concept
means learning to classify stimulus situations in terms of abstracted
properties like color, shape, position, number and others" (Gagné,
1970, p. 51). Individuals who understand a concept can do at least
four things: (1) identify the concept when examples of it are pre-
sented, (2) state properties of the concept, (3) distinguish examples
from nonexamples, and (4) better solve problems that include the
concept (Davis, Alexander, & Yelon, 1974, p. 221). However, "the
best demonstration of knowledge of the concept is the . . . ability
to differentiate examples from nonexamples" (p. 222). Put another way, "Knowing a concept involves being able to classify objects or events" (p. 221), a crucial task in any content analysis effort. It requires intellectual abilities and skills classified under application in the Taxonomy of Educational Objectives (Bloom et al., 1956, pp. 120-143). The reader is referred to Klausmeier & Harris (1966) for an anthology of papers on the topic of concept learning, and Winston (1984, chaps. 11 & 12) for a summary of computer-based, artificial intelligence implementations of concept learning.

Krippendorff's view of audits most closely resembles the quasi-evaluation approach called accountability in this chapter. His view of evaluation as a whole is characteristic of most content analysts in that it fails to distinguish between inferences based on knowledge and inferences based on value statements. This lack of distinction creates problems in justifying the standards selected. All three authors (Berelson 1952, p. 46; Holsti, 1969, p. 56; Krippendorff, 1980, p. 39) agree most content analysis studies that compare messages to a set of evaluative standards provide little or no justification for the selection of the standards used. This problem can be addressed by thinking of content analysis as a quasi-evaluation approach and using evaluation standards of quality (e.g., ERS Standards Committee, 1982; Joint Committee, 1981) when content analysis inferences produce determinations of merit or determinations of worth, rather than "simply" impartial conclusions.

Thus, content analysts have identified how messages can be compared to standards along three different dimensions. First, the
creators of the standards can be identified. The three main creators of standards are the producer of the message, the content analyst, and reference groups selected by the analyst. Second, the types of justifications for standards can be identified. Standards have been justified before the fact, because they reflect the normative characteristics of the applicable type of message, or because they are empirically derived from related, non-content sources. Third, the type of information they produce can be identified. Using the language of this study, comparisons of messages to standards grounded in concept learning produce characterizations. Comparisons of messages to standards grounded in value statements produce appraisals.

Consequences of Communications

Making inferences about the consequences of communications is a problem in pragmatics—identifying the relationships of signs (messages) to the effects they produce on the receivers of the messages. Using experimental design terminology, the message is the independent variable and the effect of the message is the dependent variable. In this case, the independent variable is measured through content analysis. The dependent variable can be measured in one of two basic ways: (1) content analysis of audience messages produced in response to the original message, or (2) any other behavioral measures that do not require content analysis.

Of course, factors besides the message of interest will also have an impact on its audience. For example, preexisting attitudes of the audience, the perceived credibility of the communicator, and
personality characteristics of the audience can all temper the effects of the message (Holsti, 1969, p. 88). Content analysis studies with adequate experimental or quasi-experimental designs (e.g., Campbell & Stanley, 1963; Cook & Campbell, 1979) can at least partially control for some of these factors. Holsti (1969) also notes, "Owing to the possible effects of factors other than message content; including audience predispositions and decoding habits, the effects of communication cannot be directly inferred from the attributes of content (what) or style (how) without independent validation" (p. 88).

One question—"with what effect?"—applies to making inferences about the consequences of communications. This question refers to the communication element of decoding the message. The uses of content analysis in this group are to: (a) measure readability, (b) analyze the flow of information from one source to another, and (c) assess responses to communications. The reader is referred to Berelson (1952, pp. 98-108) and Holsti (1969, pp. 87-93) for discussions of specific studies designed to make inferences about the consequences of communications.

Tasks of Content Analysis

Good content analysis has been defined as the high quality process of delineating, obtaining, providing, and applying descriptions and conclusions about some object; and managing and evaluating the content analysis. For any content analysis effort, the action of delineating information involves identifying the conceptual framework
for the effort; developing plans for obtaining, providing, and applying information; and developing plans for managing and evaluating the effort. Obtaining content analysis information involves the four key tasks of unitizing, sampling, coding, and analyzing messages. Providing the information involves some kind of interaction between the analyst and the audiences for the study. Applying the information depends on the intended uses for the study and any unintended uses that are also attempted. Managing the effort involves allocating and coordinating the personnel, physical resources, time, and money for all the actions of the effort. Evaluating the effort involves characterizing it and establishing its quality. Two key standards of quality for any content analysis effort are the concepts of reliability and validity. The authors of evaluation standards (ERS Standards Committee, 1982; Joint Committee, 1981) consider reliability and validity to be most clearly related to how information is delineated and obtained. Because content analysis is considered to be a quasi-evaluation approach, this relationship should also hold for such studies. As a result, the remainder of this section focuses on delineating and implementing the tasks of unitizing, sampling, coding, and analyzing messages.

**Unitizing Messages**

Because messages are very complex entities and they are produced in such high volumes, they cannot usually be properly analyzed in their entirety. To address this problem, messages can be divided into three basic types of units: (1) sampling units, (2) coding
units, and (3) context units. "Unitizing involves defining these units, separating them along their boundaries, and identifying them for subsequent analysis" (Krippendorff, 1980, p. 57). Each of these types of units are described next.

**Sampling Units.** Whenever a class of messages is so large all the messages cannot be included in the analysis, some kind of sample must be drawn. Sampling units should be constructed so that they are independent of each other. "Here 'independent' is synonymous with unrelated, unbounded, unordered, or free so that the inclusion or exclusion of any one sampling unit as a datum in an analysis has neither logical nor empirical implications for choices among other units" (Krippendorff, 1980, p. 57). Three types of sampling units are generally used: (1) sources, (2) complete documents (messages), and (3) sections of documents (Holsti, 1969, P. 130). Sources include people, government agencies, publishers, broadcasters, and other such entities. Complete "documents" include things like books, magazines, reports, interview transcripts, speeches, TV shows, movies, and photographs. Sections of documents include specific pages, editorials, responses to specific questions, time or space segments, and so on. Depending on the size of the complete set of available messages and the design of the analysis, more than one type of sampling unit can be used for any particular study.

**Coding Units.** Holsti (1969) defines a coding unit as "the specific segment of content that is characterized by placing it in a given category" (p. 116). Most coding units used in content analysis studies can be placed into one of five basic groups: "[1] words,
themes, characters, items, and space-and-time measures" (Berelson, 1952, p. 136).

The word as a coding unit is self-explanatory. It has been used most often in studies on readability, style, psychotherapy, and disputed authorship (Holsti, 1969, p. 116).

The theme is an assertion about an attitude object. "In its most compact form, the theme is a simple sentence, i.e., subject and predicate" (Berelson, 1952, p. 138). Osgood's evaluative assertion analysis (Osgood, 1959; Osgood et al., 1956) is an example of using this type of coding unit. However, because most sentences can be transformed into two or more simple assertions, Holsti (1969) argues, "This process of reducing a grammatical unit into thematic units—sometimes called 'unitizing'—can seriously reduce reliability unless the structural properties of the thematic units are precisely defined" (p. 117). Another name for transforming sentences into simple assertions or themes is "kernelizing" (Krippendorff, 1980, p. 62).

The character as a coding unit usually refers to real or fictional people. More broadly defined as referential units, this category can include "particular objects, events, persons, acts, countries, or ideas to which a particular expression refers" (Krippendorff, 1980, p. 61). This type of unit has most often been used to study the portrayals of different types of characters in different media (Holsti, 1969, p. 117).

An item is "the whole 'natural' unit employed by the producers of symbol material" (Berelson, 1952, p. 141). Examples of items include books, reports, newspaper and magazine articles, speeches,
radio and TV programs, responses to open-ended questions, and so on. Concerning this type of coding unit Berelson (1952) notes, "Analysis by the entire item is appropriate whenever the variations within the item are small or unimportant. . . . But if detailed categories are added which introduce variations within items, the item-unit may become inappropriate" (p. 141).

Like the item, space-and-time measures are examples of physical coding units. In this case, however, items are subdivided by some convenient measuring unit like the column-inch of print, minutes of broadcast time, feet of film or video tape, or grid squares over photographs. Such units have been applied almost exclusively to studies designed to identify the subject matter covered by complete items (Berelson, 1952, p. 143).

Context Units. A context unit is "the largest body of content that may be searched to characterize a recording unit" (Holsti, 1969, p. 118). The size of the context unit can be equal to or larger than the coding unit but never smaller. Perhaps the most common practice is to completely ignore context units (consciously or unconsciously), making it the same size as the coding unit. However, consciously determining the size of the context unit is crucial because its size can affect the results of the analysis. For example, Geller, Kaplan, and Lasswell (1942) demonstrated that as the size of the context unit grew from a single sentence to complete newspaper editorials, the proportion of coding units scored "neutral" diminished. As another example, for the content analysis method called contingency analysis (Osgood, 1959), in which inferences are based on the co-occurrences
of coding units within a larger context unit, the results are highly
dependent on the number of words in the context unit. However,
Osgood "found contingency values to be roughly constant between 120
and 210 words as units" (p. 62). Krippendorff (1980) underscores the
importance of context units as follows:

Context units set limits to the contextual information
that may enter the description of a recording unit. They
delineate that portion of the symbolic material that needs
to be examined in order to characterize a recording unit.
By defining a larger context unit for each recording unit,
the researcher recognizes and makes explicit the fact that
symbols codetermine their interpretation and that they
derive their meaning in part from the immediate environ­
ment in which they occur. (p. 59)

Sampling Messages

Although the sampling of messages is usually required in a
content analysis effort, the methods available are no different than
those for any other social science discipline. Because of this, only
one particularly difficult sampling problem for most content analysis
efforts is discussed. After an analyst has determined what kinds of
materials are most relevant to the study of interest, the volume of
such materials probably will still be too large to analyze complete­
ly. As a result, sampling of materials is necessary. However,
another problem is the material actually available might not truly
represent the population of interest for the study. The concern for
this issue is widely shared by content analysts (e.g., Berelson,
result, analysts often must decide if they want to draw a sample
representative of the available materials, or draw a sample representative of an underlying population differentially represented in the available materials. Krippendorff (1980, p. 68) recommends using proportional sampling techniques when making inferences about a population is desired and if it is suspected the available materials do not adequately reflect the population of interest. However, this technique does require at least a tentative hypothesis about how the population is misrepresented by the available materials.

**Coding Messages**

Holsti (1969) succinctly describes the basic process and role of coding in content analysis as follows:

> Coding is the process whereby raw data are systematically transformed and aggregated into units which permit precise description of relevant content characteristics. The rules by which this transformation is accomplished serve as the operational link between the investigator's data and his [or her] theory and hypotheses. (p. 94)

Because of the complexity of this process, good coding instructions must be explicitly developed to ensure the quality of the effort. Krippendorff (1980) suggests coding instructions should include:

- [a] a prescription of the characteristics of the observers (coder, judges) employed in the coding process;
- [b] an account of the training these observers undergo to prepare themselves for the task;
- [c] a definition of the recording [and context] units including procedures for their identification;
- [d] a delineation of the syntax and the semantics of the data language (variables, categories) including, when necessary, an outline of the cognitive procedures to be employed in placing data into categories; and
- [e] a description of how data sheets are to be used and administered. (p. 174)
These subtasks and the specifications for their instructions are discussed next.

**Observers.** Two characteristics of observers (coders) are particularly important if they are to adequately code messages (Krippendorff, 1980, p. 72). First, they must be familiar with the nature of the material to be recorded. This condition is best met when the social background of the coders is similar to the social background of the producers of the material. Second, they must be capable of reliably applying the data language (category system) to the materials. Coders with at least some background in social science methods are desirable, but all coders will need training on the specifics of the coding problem at hand.

**Training.** Because of the complexity of many content analysis efforts, the training of observers and the development of specifications for coding and context units, data languages, and data sheets is an interactive, iterative process. Although all studies will have their unique features, Krippendorff (1980) presents a set of typical activities for training observers.

[a] The research designer formulates his [or her] initial data requirements. [b] He [or she] familiarizes himself [or herself] with the way relevant information is expressed in the source material. [c] He [or she] formulates written recording instructions. [d] Working with the coders who are to apply them, instructions are jointly interpreted and modified until they meet suitable reliability requirements. . . . [e] Recording instructions are tested with a fresh set of independent observers. (pp. 73-74)

The final step is intended for studies likely to be replicated by other analysts elsewhere or by the same analysts with a different set
of materials. Without it, the effects on reliability of observers participating in the development of coding instructions cannot be known.

**Recording and Context Units.** Recording and context units have already been discussed under the task of unitizing. The type of units defined for the study depends on its theoretical framework, hypotheses, and operational transformation into a data language. When the units have easily identifiable physical boundaries (e.g., words, sentences, paragraphs, column-inches, articles, magazines, books, complete responses to questions) the specifications for identifying them are straightforward. However, when the units are primarily symbolic (e.g., themes, assertions, characters, concepts) the instructions must specify what constitutes a unit and the steps needed to identify one. If context units are not the same size as coding units, the steps needed to identify them must also be explicitly specified.

**Data Language.** The data language constitutes the set of rules and categories through which raw data is operationally linked to the analyst's theories and hypotheses. For the remainder of this paper the term, category system, is used to represent this set of rules and categories. The central role categories play in content analysis is highlighted by Berelson (1952), "Content analysis stands or falls by its categories. Particular studies have been productive to the extent that the categories were clearly formulated and well adapted to the problem and to the content" (p. 147). However, the development of productive categories is by no means a simple task, and
several attempts at identifying the appropriate categories are often required. Holsti (1969) comments on the iterative nature of the process as follows:

In the absence of standard schemes of classification, the analyst is usually faced with the task of constructing appropriate categories by trial and error methods. This process consists of moving back and forth from theory to data, testing the usefulness of tentative categories, and then modifying them in light of the data. (p. 104)

Krippendorff (1980, pp. 121-123) describes a number of computer-implemented techniques that can be used to help develop useful categories. These techniques are described in the section on performing content analysis tasks with microcomputers.


Categories must first and foremost reflect the purpose of the research, otherwise, coding the data has no meaning. Holsti (1969) elaborates as follows:

This means, first of all, that the analyst must define clearly the variables he [or she] is dealing with (the "conceptual definitions"), and secondly, he [or she] must specify the indicators which determine whether a given content datum falls within the category (the "operational definition"). A good operational definition satisfies two requirements: it is a valid representation of the analyst's concepts, and it is sufficiently precise that it guides coders to produce reliable judgments. (p. 95)
The requirement that categories be exhaustive means all coding units must be capable of being placed into a category. If this were not the case, coding units would be systematically excluded from the analysis—biasing the results.

The condition of mutually exclusive categories requires each coding unit to be placed into no more than one category. This means the operational definitions of variables must be unambiguous, and the amount of information in a coding unit must approximate the amount of information required to assign it to a single category.

The requirement for independence of assigning coding units to categories reflects one of the basic assumptions of many statistical analysis techniques. When complete messages are divided into many coding units, this assumption might be violated. In such cases, to the extent subsequent analyses rely on the assumption of independently assigned coding units, the results will be biased.

The criterion that categories should be derived from a single classification principle provides the basis for constructing mutually exclusive categories. When a group of categories reflects two or more underlying variables as if they were one, individual coding units can usually be assigned to more than one category.

For example, the four categories, positive, negative, teachers, and superintendent, reflect two variables—evaluation and staff position. Because the statement, "The teachers work their tails off every day," applies to both the positive and teachers categories, the category system violates the criterion of mutually exclusive categories. The common remedy is to "cross" the two variables (e.g.,
Kerlinger, 1986, pp. 129-130; Holsti, 1969, pp. 100-101) so that the set of categories reflects all combinations of the possible values for each variable. In the above example, the categories would become positive-teachers, positive-superintendent, negative-teachers, and negative-superintendent.

An alternative approach is to use a branching decision scheme. "Decision schemes regard each datum as the outcome of a predefined series of decisions" (Krippendorff, 1980, p. 77). He lists the advantages of decision schemes as follows:

- First, decision schemes can avoid problems arising from categories that are on different levels of generality or overlapping in meaning. . . . Second, when recording units are multidimensional, decision schemes offer the opportunity of decomposing a complex judgment into several simple decisions and thereby achieve levels of reliability not obtainable otherwise. Third, decision schemes drastically reduce the number of alternatives to be simultaneously considered at each step. (pp. 77-78)

For the above variables, evaluation and staff position, if we are interested in focusing in on the targets of negative evaluations, but we do not care about the targets of positive evaluations, we could use a two-step decision scheme as follows: (1) divide the evaluative statements into positive and negative groups, then (2) divide the negative statements into those about teachers and those about the superintendent. This decision scheme requires only three categories (positive, negative-teachers, and negative-superintendent) and it meets the criterion of mutually exclusive categories. Thus the statement, "The teachers work their tails off every day," would be unambiguously placed into the positive category. Examples of many
other kinds of category systems are summarized in a number of sources (e.g., Berelson, 1952, pp. 147-168; Dunphy, 1966; Holsti, 1969, pp. 95-116; Krippendorff, 1980, pp. 75-81, 85-118).

Data Sheets. Data sheets are the medium on which the coding process is documented. Because of this, they must contain all the information necessary to represent the placement of coding units into categories, as well as some "housekeeping" information. Krippendorff (1980) suggests data sheets should contain at least three types of information: "[1] administrative information, [2] information on data organization, and [3] information on the phenomena to be recorded, the data" (p. 82). His examples of administrative information include: (a) identification of the project to which the data apply, (b) identification of the stage of processing of the data, (c) identification of the individuals who processed the data at each stage, and (d) instructions on how the data should be transferred to computer data files. Information on data organization refers to how data on a particular sheet relates to data on other sheets. For example, demographic information on data sources is usually kept on one type of sheet and the messages they produce are usually recorded on another type of sheet. Unique identification numbers for the sources placed on both sheets maintain the relation between the sources and their messages. Finally, the data (coding units) are on the sheets. Depending on the size of the context units, they might also be on the same sheets. The representation of the data and possible categories into which they can be coded depends on the exact nature of the data and the category system used. Because references
to examples of different category systems have already been provided, they are not repeated here.

Analyzing Messages

As reflected in the first criterion for constructing good categories, the nature of the analysis of a set of messages should be intimately related to the purposes of the effort. The previous discussion on the uses of content analysis summarized three general purposes for conducting a content analysis. They were to: (1) make inferences about the antecedents of communications, (2) describe and make inferences about the characteristics of communications, and (3) make inferences about the consequences of communications. Holsti (1969, pp. 27-37) presents a series of general designs that can be use to organize content analysis efforts directed toward each of these purposes. The designs of actual content analysis studies should reflect their general purpose and the specific theories, hypotheses, and variables that apply. Because of the interdependent nature of designing category systems and analytical procedures, the references listed for examples of category systems also contain many examples of specific analytical procedures.

Performing Content Analysis Tasks with Microcomputers

Microcomputer programs can be of use for many kinds of administrative tasks. Because of this, most research and evaluation organizations are likely to already have a battery of microcomputers and general purpose programs available to them. Gray (1984d) lists
five types of "administrative" or management programs of use to the
graphics" (p. 80). Two types of programs—word processing and data
base management—are particularly relevant to this discussion. The
reasons for this are discussed later in this section. Elsewhere,
Gray presents more detailed discussions of how word processing pro-
grams (1984a) and data base management programs (1984b, 1984c) can be
used for administrative purposes in the field of evaluation.

These general uses for microcomputer programs are important
because they provide the basic justification for purchasing the
microcomputers and programs in the first place. However, once they
are available to the organization, these resources can also be adapt­
ed to content analysis uses.

Besides general project management and support activities like
word processing and data base management, computers can be put to
three general uses in content analysis efforts (Krippendorff, 1980). They include statistical analyses, computational aids for survey and
discovery, and computational content analysis (pp. 119-128).

Statistical analyses are not unique to content analysis efforts
and they are not of particular interest here. Common descriptive and
inferential statistics familiar to social scientists in general are
also of use in many content analysis studies.

Computational aids for survey and discovery help content
analysts consolidate large masses of textual material so that various
types of overviews of the information contained in them can be
developed. Such overviews can be used during the development of coding instructions to help identify applicable categories. Similar techniques can also be used to help place sets of coding units into existing categories. In computational survey and discovery, the human still makes all the "hard decisions" and simply uses the computer to perform a number of "clerical" functions. This is the use of computers of interest for this study.

Computational content analyses are performed primarily by computer programs rather than by humans. Such programs are simultaneously very complex and overly simplistic. That is, the programs themselves are very large and complicated, requiring high powered mainframe or supermini computers; while their performance is usually narrowly focused and often lacking the "common sense" of even a novice content analyst. The best example of this high powered type of program is the General Inquirer (Stone et al., 1966).

While Krippendorff and other authors (e.g., Gerbner, Holsti, Krippendorff, Paisley, & Stone, 1969; Holsti, 1969, pp. 150-194) discuss a number of variations of computer-assisted content analysis, the techniques most useful for survey and discovery can be placed into three basic groups: (1) key words out of context, (2) key words in context, and (3) information retrieval. All these techniques can be implemented with custom-designed computer programs running on large or small computers. However, some of them can also be implemented on programs originally designed for other purposes. Variations of each technique and some of the general purpose programs that can be used to implement them are discussed next.
Key words out of context are basically word lists. The lists are usually of single words but they can also be of phrases or groups of words that occur within a specified distance of each other (e.g., no more than five words apart). The frequency of occurrence of each item in the document is also listed. The list can be ordered alphabetically or by frequency of occurrence. Finally, items with high, low, or chance frequencies of occurrence; or types of words like articles, prepositions, and pronouns; can be deleted from the list.

Word lists are relatively easy to produce for a skilled computer programmer with just about any programming language, such as BASIC or Pascal. However, because of the way some "spelling checker" programs are designed, they automatically produce word lists. If these lists are accessible to the user, they can also be considered key word out of context lists. One such program for microcomputers is called The Word Plus (Holder, 1982). An option of this program is to create a text file that lists all the unique words contained in a different text file (p. 38). The number of times each word appears in the source file (e.g., a collection of coding items) is also included in the list. The list can be ordered alphabetically or by frequency of occurrence. Because this list is a text file, it can be edited with a word processing program. This means unwanted words like articles, pronouns, or those with low frequencies can be easily removed from the list. The list can then be used to help decide what categories should be used in the final content analysis.

Key words in context are lists of specified words surrounded by parts of the text in which they occur. This shows the reader how the
word was used in context. The length of the text is usually short enough to be printed on one line with the key word centered. The line can also be indexed so the source material is easily accessible. This is a very special type of list that is more difficult for a programmer to produce than simple word lists. In addition, no computer programs designed for general business or educational uses produce this kind of list. Wood (1984) summarizes several qualitative and quantitative social research uses for this type of list.

The third type of technique, information retrieval, can be used on "original" documents, such as complete word processing files, or textual data base files in which each "record" can contain one coding unit identified from a larger document. The two most common information retrieval functions that can be performed on these files are searching for and sorting information. Once found or sorted, the information can then be displayed to the user in any number of ways.

When the material is basically "free form," like the chapter of a book, information retrieval is primarily limited to searching for and displaying specified words or phrases. Just about any word processing program has this capability, although the results usually can only be presented on the screen. Depending on the particular word processing program, a section of text containing the specified items could be "cut" from the document and then "pasted" into a different document with similar passages, but the following approach is much more powerful and convenient overall.

When the text (recording unit) is organized into a database as one "field" within a larger "record," both searching and sorting can
take place on any one or a combination of fields. This allows for very flexible and powerful manipulations of the textual material with relatively little effort on the part of the user, particularly when the other fields of a record contain relevant information about the textual material. A further advantage of this approach is the results of searches and sorts can be sent to a number of destinations, such as the screen, printers, and other files. Examples of both of these capabilities in a microcomputer word processing program are WPS List Processing (Digital Equipment Corporation, 1984a) and WPS Sort (Digital Equipment Corporation, 1984b). An example of a microcomputer data base management program with searching, sorting, and displaying capabilities is dBASE II (Ratliff, 1982). This technique can be used during the process of developing a category system or while coding units for an existing set of categories.

Many evaluation and research oriented organizations now have microcomputers with a number of general purpose, business application programs like those for word processing and data base management. As a result, they also already have a basic library of programs that can be adapted to many survey and discovery uses in content analysis efforts. The knowledge of a few simple techniques and a lot of imagination are the keys to discovering these uses.

Summary

This chapter was used to present the conceptual and operational relationships between evaluation, content analysis, and microcomputers. The discussion was divided into four major topics: (1) the
conceptual relationships between evaluation and content analysis, (2) an overview of evaluation approaches, (3) a general model for conducting an evaluation effort, and (4) an overview of content analysis.

The key concepts for identifying the relationships between evaluation and content analysis were organized into three basic groups: (1) information, (2) actions, and (3) standards of quality. The relationships between these concepts and the two fields were summarized by presenting working definitions of evaluation and content analysis that used the concepts and conformed to a common grammatical structure. Both definitions are repeated here. Good evaluation is the high quality process of delineating, obtaining, providing, and applying characterizations and appraisals about some object; and managing and evaluating the evaluation. Good content analysis is the high quality process of delineating, obtaining, providing, and applying descriptions and conclusions about some object; and managing and evaluating the content analysis. These definitions contain direct references to the information and action components that were identified. However, because of the emerging status of standards of quality in both fields, only indirect references were made to them in the definitions. In any event, the two standards of quality currently found in common between the two fields are those of reliability and validity.

The overview of evaluation approaches was organized around a classification framework with three dimensions: (1) epistemology, (2) major perspective, and (3) orientation to values. Although this
framework contained twelve cells, fifteen general evaluation approaches were found to fill only seven of them. Within this framework, content analysis was found to be an objectivist, elite, quasi-evaluation approach. Each of the evaluation approaches were further summarized in terms of their main organizer, purpose, strengths, and weaknesses.

The general model for conducting an evaluation effort provided a graphic representation of the working definition of evaluation. As such, it represented the key components and their logical relationships for an evaluation effort. A decision network based on the general model was also presented. It represented the dynamic and iterative nature of processing evaluation information. Because content analysis was classified as a quasi-evaluation approach, this model also applies to content analysis efforts.

The overview of content analysis contained four main discussions: (1) the classification of content analysis uses, (2) a summary of content analysis uses, (3) a presentation of the tasks crucial to delineating and obtaining high quality content analysis information, and (4) the ways microcomputers can be used to help delineate and obtain content analysis information. Fifteen content analysis uses were classified in relation to the basic components of the communication paradigm. This paradigm can be represented by the compound question, "Who, said what, to whom, why, how, and with what effect?" In the summary, groups of content analysis uses were discussed in terms of their general purposes and the communication question they addressed. Specific uses were discussed to the extent
they could be related to the field of evaluation. One such use was to make inferences about the attitudes of a selected group of people based on the messages they produced. A common way to collect such messages in evaluation efforts is to ask them to respond to a set of open-ended survey questions.

Four key tasks involved in obtaining content analysis information for any purpose were further described. These tasks included unitizing, sampling, coding, and analyzing messages. Finally, three ways of using computers in content analysis tasks were identified. One of the ways, as computational aids for survey and discovery, was considered to be the most promising avenue for using microcomputers to help obtain content analysis information, particularly when developing coding instructions or actually coding messages. Two survey and discovery techniques that can be implemented using certain kinds of microcomputer programs not specifically designed for content analysis were identified. They included: (1) producing key words out of context lists with certain kinds of spelling checker programs, and (2) performing information retrieval activities such as searching for, sorting, and displaying messages with certain kinds of word processing programs or data base management programs.
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.** $E_1$ = Experiment 1, $E_2$ = Experiment 2, $T_0$ = Pre-assignment training, $R$ = Random assignment to groups, $T_1$ = $E_1$ experimental group treatment, $T_2$ = $E_1$ control group treatment, $O_1$ = $E_1$ observation, $T_3$ = $E_2$ experimental group treatment, $T_4$ = $E_2$ control group treatment, $O_2$ = $E_2$ 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.

<table>
<thead>
<tr>
<th>Category experiment design</th>
</tr>
</thead>
<tbody>
<tr>
<td>T0 R</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experimental procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coding experiment design</th>
</tr>
</thead>
<tbody>
<tr>
<td>T0 R</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Legend. T = Treatment n, R = Random assignment, O = Observation n, S = Session n, Ta = Task n, C = 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 \( \frac{1}{\text{the number of categories established by the Hierarchy Panel}} \), or \( \frac{1}{19} \).

**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 $1/5$ ($1/n$ 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.
CHAPTER IV

RESULTS

Introduction

This chapter is used to report the results of the two experiments described in Chapter 3. They were designed to address the following problem: How can evaluation practitioners use microcomputer programs to obtain reliable and valid content analyses of responses to open-ended survey questions?

The procedures of the experiments were organized into a single simulation activity in which participants were asked to assume the role of a student assistant at a university research center. In this role, the student was asked to summarize a set of responses to an open-ended question used in an evaluation project. This assignment was divided into four tasks: (1) develop a category system, (2) verify the category system, (3) code the set of responses in terms of the final category system, and (4) verify the codes. Both experiments began during a classroom lecture used to introduce the study. Participants completed each task out of class and exchanged materials with the researcher at the beginning of subsequent classes. Experiment 1 ended after the second task. Measures of category reliability—operationalized as agreement—and validity were recorded at that time. Experiment 2 ended after the fourth task. Measures of coding reliability and validity were recorded after this final task.

133
Three main topics are discussed in the remainder of this chapter: (1) the results of Experiment 1, (2) the results of Experiment 2, and (3) comparisons of those who did or did not complete each experiment. For Experiments 1 and 2, the independent variable, research hypotheses, dependent variables, and null hypotheses are reviewed before the results of the data analyses are presented. For the comparisons of those who did or did not complete each experiment, a summary of selected characteristics is presented, followed by a summary of tests to determine if any observed differences between the groups are statistically significant.

Effects of Microcomputer Output on Category Development

Experiment 1 was used to test the effects of specialized microcomputer output on the reliability and validity of developing a set of content analysis categories. These categories were used to classify a set of responses to an open-ended survey question used in the simulated evaluation effort. Seventy-four students from six College of Education classes completed Experiment 1.

The independent variable was the possession or lack of possession of two types of microcomputer output: (1) a word count list sorted by frequency of occurrence derived from the complete set of responses used in the simulation, and (2) half of the responses sorted according to how they were coded by each participant, and labeled with participant-developed category identifiers and summaries. Forty-two randomly-assigned experimental participants received word count lists at the beginning of the first task while control
participants never received such lists. Experimental participants received responses sorted and labeled according to their own categories at the beginning of the second task. Control participants (32) received unsorted and unlabeled responses at that time.

The first and second research hypotheses of the study were identified for Experiment 1. They were:

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.

The corresponding dependent variable, null hypothesis, and results of analyses for category reliability are discussed next. This is followed by a comparable discussion for category validity.

**Category Reliability**

Category reliability was defined as the extent to which the same set of categories was created from the simulation documents and responses used in the category experiment under varying circumstances at different locations, by different participants. The median proportion of agreement was used as the measure of category reliability.
for each participant because: (a) it reflects the extent to which each participant agreed with the rest of his or her treatment group about which five categories should be included in the response classification system, and (b) analysis of variance can be used to test the null hypothesis related to group mean scores on category reliability when it is operationalized as a proportion of agreement.

However, the proportion of agreement between any two participants that is possible by chance alone is greater than zero \( \frac{1}{\text{number of categories established by the Hierarchy Panel, or 1/19}} \). Thus, the aggregated median score for each individual, and the mean score for each group also have this expected value. When both group mean scores could have been obtained by chance alone, testing the difference between these means would be pointless, even when the obtained scores are greater than zero.

This possibility was checked by generating and testing the applicable null hypotheses related to the mean scores for each group. They are: (a) no difference exists between the mean of the median proportion of category agreement scores for the experimental group and the mean of such scores that would be expected by chance alone, and (b) no difference exists between the mean of the median proportion of category agreement scores for the control group and the mean of such scores that would be expected by chance alone. These null hypotheses were tested using a one-sample t-test in which the observed group mean was compared to the group mean expected by chance.

The null hypothesis relevant to research hypothesis 1, category reliability, is as follows: no difference exists between the mean of
the median proportion of category agreement scores for the participants who received specialized microcomputer output and the mean of such scores for the participants who received no specialized microcomputer output. A one-way analysis of variance was used to test this null hypothesis.

The group means and standard deviations for category reliability are presented in Table 9. The results of the t-tests are presented in Table 10. The results of the analysis of variance are presented in Table 11.

Table 9

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Observed mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>42</td>
<td>0.47</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>0.42</td>
<td>0.16</td>
<td></td>
</tr>
</tbody>
</table>

Table 10

Summary of t-Tests Between Observed Mean Scores and Corresponding Expected Scores Due to Chance Alone for Hypothesis 1, Category Reliability

<table>
<thead>
<tr>
<th>Critical comparison</th>
<th>Observed</th>
<th>Chance</th>
<th>t</th>
<th>Critical value</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>0.47 to 0.053</td>
<td>16.89</td>
<td>2.02</td>
<td>Reject</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>0.42 to 0.053</td>
<td>12.88</td>
<td>2.04</td>
<td>Reject</td>
<td></td>
</tr>
</tbody>
</table>
Table 11

One-Way Analysis of Variance for Hypothesis 1, Category Reliability

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F</th>
<th>Sig. of F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>0.050</td>
<td>1</td>
<td>0.050</td>
<td>1.957</td>
<td>0.166</td>
</tr>
<tr>
<td>Within</td>
<td>1.854</td>
<td>72</td>
<td>0.026</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1.905</td>
<td>73</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As indicated in Table 9, the observed mean proportion score for the experimental group (0.47). The observed mean proportion score for the control group (0.42). In addition, as shown in Table 10, both of these observed scores are higher than would be expected by chance alone (p < 0.05). Thus, the corresponding null hypotheses were both rejected. Finally, the analysis of variance summarized in Table 11 indicates the difference between the two observed mean scores is not greater than would be expected by chance alone (p < 0.10). Thus, the null hypothesis stating no difference exists between the experimental group mean and the control group mean on category reliability was retained.

Category Validity

Category validity was defined as the extent to which the set of categories created by participants in the category experiment agreed with the set of categories created by the Response Panel. The total number of agreements with the Response Panel was selected as the measure of category validity for each participant.
The null hypothesis relevant to research hypothesis 2, category validity, is as follows: no difference exists between the mean total number of category agreements with the Response Panel for the participants who received specialized microcomputer output and the mean of such scores for the participants who received no specialized microcomputer output. A one-way analysis of variance was used to test this null hypothesis.

The results of the analysis of variance identified above are presented in Table 12. The corresponding group means and standard deviations are presented in Table 13.

Table 12
One-Way Analysis of Variance for Hypothesis 2, Category Validity

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F</th>
<th>Sig. of F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>2.474</td>
<td>1</td>
<td>2.474</td>
<td>1.479</td>
<td>0.228</td>
</tr>
<tr>
<td>Within</td>
<td>120.405</td>
<td>-72</td>
<td>1.672</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>122.878</td>
<td>73</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 13
Group Means and Standard Deviations for Hypothesis 2, Category Validity

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Observed mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>42</td>
<td>3.12</td>
<td>1.21</td>
</tr>
<tr>
<td>Control</td>
<td>32</td>
<td>2.75</td>
<td>1.39</td>
</tr>
</tbody>
</table>
As summarized in Table 12, no statistically significant difference \( (p \leq 0.10) \) in mean category validity scores was found between those groups of participants who did or did not receive specialized microcomputer outputs. Therefore, the null hypothesis was retained. Thus, the difference between the obtained mean score for the experimental group (3.12) and the obtained mean score for the control group (2.75) is considered to have occurred by chance alone.

Effects of Microcomputer Output on Response Coding

Experiment 2 was used to test the effects of specialized microcomputer output on the reliability and validity of coding a group of responses to the simulation's open-ended question. These codes were based on the final set of content analysis categories established by the Response Panel. Fifty-nine students completed Experiment 2.

Just as in Experiment 1, the independent variable was the possession or lack of possession of two types of microcomputer output: (1) a word count list, and (2) responses sorted according to how they were coded by each participant, and labeled by category identifiers and summaries. The design of Experiment 2 subsumed both of the participant tasks of Experiment 1 plus two additional tasks. As a result, the discussion of the Experiment 1 materials supplied to participants for Task 1 and Task 2 applies here as well.

For Task 3, experimental participants (30) received the 50 original responses sorted and labeled according to their own categories, as updated during the previous task, plus 50 new responses at the end of the list. For Task 4, the experimental participants
received all 100 responses sorted and coded according to the final set of categories used by all participants. As usual, control participants (29) received unsorted and unlabeled responses for Task 3 and Task 4. For both tasks, they received the original 50 responses plus 50 additional responses in the same order, differing only in code changes made by each participant during the previous task.

The third and fourth research hypotheses of the study were identified for Experiment 2. They were:

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.

The corresponding dependent variable, null hypothesis, and results of analyses for coding reliability are discussed next. This is followed by a comparable discussion for coding validity.

Coding Reliability

Coding reliability was defined as the extent to which the same codes were assigned to the coding experiment responses under varying circumstances, at different locations, by different participants.
The median proportion of agreement was used as the measure of coding reliability for each participant because: (a) it reflects the extent to which each participant agreed with his or her treatment group about how the 100 responses should be coded, and (b) analysis of variance can be used to test the null hypothesis related to group mean scores when coding reliability is operationalized as a proportion.

The expected value for the proportion of agreement between any two participants due to chance alone is \( \frac{1}{5} \) (1 / number of final categories used by all participants). The aggregated median score for each participant, and the mean score for each group also have this expected value when all agreements are due to chance alone. The null hypotheses related to testing the observed mean scores for each group against the mean scores expected by chance alone are as follows: (a) no difference exists between the mean of the median proportion of coding agreement scores for the experimental group and the mean of such scores that would be expected by chance alone, and (b) no difference exists between the mean of the median proportion of coding agreement scores for the control group and the mean of such scores that would be expected by chance alone. These null hypotheses were tested using a one-sample t-test in which the observed group mean was compared to the group mean expected by chance alone.

The null hypothesis relevant to research hypothesis 3, coding reliability, is as follows: no difference exists between the mean of the median proportion of coding agreement scores for the participants who received specialized microcomputer output and the mean of such scores for the participants who received no specialized microcomputer output.
output. A one-way analysis of variance was used to test this null hypothesis.

The group means and standard deviations for coding reliability are presented in Table 14. The results of the t-tests are presented in Table 15. The results of the analysis of variance are presented in Table 16.

Table 14

Group Means and Standard Deviations for Hypothesis 3, Coding Reliability

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Observed mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>30</td>
<td>0.82</td>
<td>0.05</td>
</tr>
<tr>
<td>Control</td>
<td>29</td>
<td>0.75</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Table 15

Summary of t-Tests Between Observed Mean Scores and Corresponding Expected Scores Due to Chance Alone for Hypothesis 3, Coding Reliability

<table>
<thead>
<tr>
<th>Critical comparison</th>
<th>Observed</th>
<th>Chance</th>
<th>t</th>
<th>Critical value</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>0.82 to 0.20</td>
<td>67.92</td>
<td>2.04</td>
<td>Reject</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>0.75 to 0.20</td>
<td>32.91</td>
<td>2.05</td>
<td>Reject</td>
<td></td>
</tr>
</tbody>
</table>

As shown in Table 14, the observed mean score for the experimental group is 0.82. The observed mean score for the control group is 0.75. As indicated in Table 15, both of these observed scores are higher than would be expected by chance alone (p < 0.05). Thus, the
corresponding null hypotheses were rejected. Finally, the analysis of variance summarized in Table 16 indicates the difference between the two observed mean scores is greater than chance \((p \leq 0.001)\). As a result, the null hypothesis stating no difference exists between the experimental and control group means on category reliability was rejected. Thus, the experimental proportion of agreement (0.82) is higher than the control proportion of agreement (0.75).

Table 16
One-Way Analysis of Variance for Hypothesis 3, Coding Reliability

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F</th>
<th>Sig. of F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>0.063</td>
<td>1</td>
<td>0.063</td>
<td>12.838</td>
<td>0.001 **</td>
</tr>
<tr>
<td>Within</td>
<td>0.280</td>
<td>57</td>
<td>0.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0.343</td>
<td>58</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** Significant at 0.001 level

Coding Validity

Coding validity was defined as the extent to which the codes assigned to the responses by participants in the coding experiment agreed with the the codes assigned to those responses by the Response Panel. The total number of agreements with the Response Panel was selected as the measure of coding validity for each participant.

The null hypothesis relevant to research hypothesis 4, coding validity, is as follows: no difference exists between the mean total number of coding agreements with the Response Panel for the
participants who received specialized microcomputer output and the mean of such scores for the participants who received no specialized microcomputer output. A one-way analysis of variance was used to test this null hypothesis.

The results of the analysis of variance noted above are presented in Table 17. The corresponding group means and standard deviations are presented in Table 18.

Table 17

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F</th>
<th>Sig. of F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>307.205</td>
<td>1</td>
<td>307.205</td>
<td>3.751</td>
<td>0.058 *</td>
</tr>
<tr>
<td>Within</td>
<td>4668.829</td>
<td>57</td>
<td>81.909</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4976.034</td>
<td>58</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Significant at 0.10 level

Table 18

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Observed mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>30</td>
<td>86.63</td>
<td>7.46</td>
</tr>
<tr>
<td>Control</td>
<td>29</td>
<td>82.07</td>
<td>10.44</td>
</tr>
</tbody>
</table>

As summarized in Table 17, the observed difference between mean coding validity scores for the group who did and the group who did not receive specialized microcomputer output is statistically
significant \((p < 0.10)\). Therefore, the null hypothesis was rejected. As noted in Table 18, the obtained mean score for the experimental group (86.63) is higher than the obtained mean score for the control group (82.07). This difference is considered to be greater than what would have occurred by chance alone, if the null hypothesis is true.

Comparisons of Those Who Did or Did Not Complete Each Experiment

Not all students enrolled in the six study classes completed the two experiments. Of the 125 people included in this study group, 78 initially consented to participate. However, four students dropped out of the study before they completed Experiment 1, and 14 others dropped out before they completed Experiment 2. In addition, one participant was dropped from Experiment 2 by the researcher for failure to properly complete a number of tasks.

Because of a low participation rate during the pilot study, certain changes were made to the original procedures. One change was to collect a set of demographic information from the students during the first class session whether they participated in the study or not. A summary of this information along with treatment group and class membership is presented in Table 19. The characteristics are summarized for all students, those who did or did not complete Experiment 1, and those who did or did not complete Experiment 2. In each cell, the total number of students with each characteristic and its corresponding percent of all 125 students are presented. The total pool of students is used as a base because they all participated in Session 1. The characteristics include: (a) the total number of
### Table 19
Characteristics of Students Enrolled in the Six Study Classes

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>All students</th>
<th>Completed experiment 1</th>
<th>Completed experiment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>N (%)</td>
<td>N (%)</td>
</tr>
<tr>
<td>Total number</td>
<td>125 100.0</td>
<td>74 59.2</td>
<td>51 40.8</td>
</tr>
<tr>
<td></td>
<td>59 47.2</td>
<td>66 52.8</td>
<td></td>
</tr>
<tr>
<td>Treatment group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>65 52.0</td>
<td>42 33.6</td>
<td>23 18.4</td>
</tr>
<tr>
<td>Control</td>
<td>60 48.0</td>
<td>32 25.6</td>
<td>28 22.4</td>
</tr>
<tr>
<td>Class</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ED 322a</td>
<td>30 24.0</td>
<td>13 10.4</td>
<td>17 13.6</td>
</tr>
<tr>
<td>ED 322b</td>
<td>28 22.4</td>
<td>17 13.6</td>
<td>11 8.8</td>
</tr>
<tr>
<td>ED 450/455</td>
<td>13 10.4</td>
<td>12 9.6</td>
<td>1 0.8</td>
</tr>
<tr>
<td>ED 516</td>
<td>24 19.2</td>
<td>12 9.6</td>
<td>12 9.6</td>
</tr>
<tr>
<td>ED 602</td>
<td>14 11.2</td>
<td>6 4.8</td>
<td>8 6.4</td>
</tr>
<tr>
<td>EDLD 663</td>
<td>16 12.8</td>
<td>14 11.2</td>
<td>2 1.6</td>
</tr>
<tr>
<td>Employed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full time</td>
<td>49 39.2</td>
<td>30 24.0</td>
<td>19 15.2</td>
</tr>
<tr>
<td>Part time</td>
<td>45 36.0</td>
<td>29 23.2</td>
<td>16 12.8</td>
</tr>
<tr>
<td>Not employed</td>
<td>28 22.4</td>
<td>15 12.0</td>
<td>13 10.4</td>
</tr>
<tr>
<td>Missing</td>
<td>3 2.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degree program</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Undergraduate</td>
<td>74 59.2</td>
<td>42 33.6</td>
<td>32 25.6</td>
</tr>
<tr>
<td>Graduate</td>
<td>48 38.2</td>
<td>32 25.6</td>
<td>16 12.8</td>
</tr>
<tr>
<td>Missing</td>
<td>3 2.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>College of Major</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arts &amp; sciences</td>
<td>40 32.0</td>
<td>22 17.6</td>
<td>18 14.4</td>
</tr>
<tr>
<td>Bus/Engr/Fine</td>
<td>13 10.4</td>
<td>6 4.8</td>
<td>7 5.6</td>
</tr>
<tr>
<td>Education</td>
<td>67 53.6</td>
<td>44 35.2</td>
<td>23 18.4</td>
</tr>
<tr>
<td>Missing</td>
<td>5 4.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Table 19 (continued)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>All students</th>
<th>Completed experiment 1</th>
<th>Completed experiment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td>Expected grade</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>64</td>
<td>51.2</td>
<td>44</td>
</tr>
<tr>
<td>BA</td>
<td>22</td>
<td>17.6</td>
<td>12</td>
</tr>
<tr>
<td>B</td>
<td>22</td>
<td>17.6</td>
<td>11</td>
</tr>
<tr>
<td>Missing</td>
<td>17</td>
<td>13.6</td>
<td>11</td>
</tr>
<tr>
<td>Proportion of total hours completed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00 to 0.40</td>
<td>37</td>
<td>29.6</td>
<td>24</td>
</tr>
<tr>
<td>0.41 to 0.80</td>
<td>34</td>
<td>27.2</td>
<td>15</td>
</tr>
<tr>
<td>0.81 to 1.00</td>
<td>15</td>
<td>12.0</td>
<td>11</td>
</tr>
<tr>
<td>1.01 to 1.40+</td>
<td>5</td>
<td>4.0</td>
<td>3</td>
</tr>
<tr>
<td>Missing</td>
<td>34</td>
<td>27.2</td>
<td></td>
</tr>
<tr>
<td>Proportion of full time enrollment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00 to 0.40</td>
<td>47</td>
<td>37.6</td>
<td>33</td>
</tr>
<tr>
<td>0.41 to 0.80</td>
<td>24</td>
<td>19.2</td>
<td>14</td>
</tr>
<tr>
<td>0.81 to 1.00</td>
<td>31</td>
<td>24.8</td>
<td>15</td>
</tr>
<tr>
<td>1.01 to 1.40+</td>
<td>19</td>
<td>15.2</td>
<td>12</td>
</tr>
<tr>
<td>Missing</td>
<td>4</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>Content analysis studies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>117</td>
<td>93.6</td>
<td>70</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>4.0</td>
<td>4</td>
</tr>
<tr>
<td>Missing</td>
<td>3</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>Other studies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>114</td>
<td>91.2</td>
<td>69</td>
</tr>
<tr>
<td>1 to 4</td>
<td>8</td>
<td>6.4</td>
<td>5</td>
</tr>
<tr>
<td>Missing</td>
<td>3</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>Total studies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>111</td>
<td>88.8</td>
<td>67</td>
</tr>
<tr>
<td>1 to 4</td>
<td>11</td>
<td>8.8</td>
<td>7</td>
</tr>
<tr>
<td>Missing</td>
<td>3</td>
<td>2.4</td>
<td></td>
</tr>
</tbody>
</table>
students, (b) treatment group membership, (c) enrolled class, (d) employment status, (e) current degree program, (f) college of major, (g) expected grade for the class, (h) the proportion of the total hours completed for the applicable degree program, (i) the proportion of full time enrollment for the degree program, (j) the number of content analysis studies in which the student participated, (k) the number of other studies in which the student participated, and (l) the total number of studies in which the student participated.

To determine if the groups of those who did or did not complete either experiment differed significantly on any of the above characteristics, the differences between total group membership for each characteristic were statistically tested at the 0.05 alpha level. Two-sample chi-square tests of differences between total group membership were computed for each of the categorical variables. One-way analysis of variance tests of differences between total group membership were computed for each of the ratio level variables.

A summary of the chi-square tests for Experiment 1 is presented in Table 20. A summary of the analysis of variance tests for Experiment 1 is presented in Table 21. Only one of the tests indicates statistically significant differences between those who did or did not complete Experiment 1. That test was on class membership. Inspection of Table 19 offers some clues to why the null hypothesis for this characteristic was rejected. Two classes, ED 450/455 and EDLD 663, had a very high participation rate—nearly 90% for the two classes combined. The remaining classes had substantially lower participation rates—50% for the four classes combined.
Table 20
Summary of Chi-Square Tests of Differences on Selected Variables Between Those Who Did or Did Not Complete Experiment 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Treatment</th>
<th>Class</th>
<th>Employment</th>
<th>Degree</th>
<th>College</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>1.210</td>
<td>16.747</td>
<td>0.866</td>
<td>0.819</td>
<td>2.353</td>
</tr>
<tr>
<td>Degrees of freedom</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Significance</td>
<td>0.271</td>
<td>0.005</td>
<td>0.649</td>
<td>0.366</td>
<td>0.308</td>
</tr>
<tr>
<td>Decision</td>
<td>Retain</td>
<td>Reject</td>
<td>Retain</td>
<td>Retain</td>
<td>Retain</td>
</tr>
</tbody>
</table>

Table 21
Summary of Analysis of Variance Tests of Differences on Selected Variables Between Those Who Did or Did Not Complete Experiment 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Expected grade</th>
<th>Proportion hrs. completed</th>
<th>Proportion FTE hours</th>
<th>Number C.A. studies</th>
<th>Number other studies</th>
<th>Total number studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>1.365</td>
<td>0.530</td>
<td>0.962</td>
<td>0.810</td>
<td>0.183</td>
<td>0.447</td>
</tr>
<tr>
<td>Sig.</td>
<td>0.246</td>
<td>0.469</td>
<td>0.330</td>
<td>0.370</td>
<td>0.670</td>
<td>0.505</td>
</tr>
<tr>
<td>Decision</td>
<td>Retain</td>
<td>Retain</td>
<td>Retain</td>
<td>Retain</td>
<td>Retain</td>
<td>Retain</td>
</tr>
</tbody>
</table>

A summary of the chi-square tests for Experiment 2 is presented in Table 22. A summary of the analysis of variance tests for Experiment 2 is presented in Table 23. Two of the tests performed indicate statistically significant differences between those who did or did not complete Experiment 2. Those tests were for class membership and degree program. Inspection of Table 19 reveals the two classes,
Table 22

Summary of Chi-Square Tests of Differences on Selected Variables Between Those Who Did or Did Not Complete Experiment 2

<table>
<thead>
<tr>
<th>Variable</th>
<th>Treatment</th>
<th>Class</th>
<th>Employment</th>
<th>Degree</th>
<th>College</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>0.004</td>
<td>24.688</td>
<td>0.218</td>
<td>5.436</td>
<td>2.903</td>
</tr>
<tr>
<td>Degrees of freedom</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Significance</td>
<td>0.949</td>
<td>0.000</td>
<td>0.897</td>
<td>0.020</td>
<td>0.234</td>
</tr>
<tr>
<td>Decision</td>
<td>Retain</td>
<td>Reject</td>
<td>Retain</td>
<td>Reject</td>
<td>Retain</td>
</tr>
</tbody>
</table>

Table 23

Summary of Analysis of Variance Tests of Differences on Selected Variables Between Those Who Did or Did Not Complete Experiment 2

<table>
<thead>
<tr>
<th>Variable</th>
<th>Expected grade</th>
<th>Proportion hrs. completed</th>
<th>Proportion FTE hours</th>
<th>Number C.A. studies</th>
<th>Number other studies</th>
<th>Total number studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>0.142</td>
<td>0.514</td>
<td>0.005</td>
<td>2.091</td>
<td>0.166</td>
<td>0.676</td>
</tr>
<tr>
<td>Sig.</td>
<td>0.707</td>
<td>0.476</td>
<td>0.944</td>
<td>0.151</td>
<td>0.685</td>
<td>0.413</td>
</tr>
<tr>
<td>Decision</td>
<td>Retain</td>
<td>Retain</td>
<td>Retain</td>
<td>Retain</td>
<td>Retain</td>
<td>Retain</td>
</tr>
</tbody>
</table>

ED 450/455 and EDLD 663 still had a very high combined participation rate, about 86%. On the other hand, the combined participation rate for the remaining four classes dropped to just over 35%. Finally, less than 40% of all undergraduates completed Experiment 2, while about 62% of all graduate students completed Experiment 2.
Summary

The results of the four data analyses related to the primary study hypotheses are summarized in Table 24.

Table 24
Summary of Analysis of Variance Tests for the Four Study Hypotheses

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>1.957</td>
<td>1.479</td>
<td>12.838</td>
<td>3.751</td>
</tr>
<tr>
<td>Sig.</td>
<td>0.166</td>
<td>0.228</td>
<td>0.001</td>
<td>0.058</td>
</tr>
<tr>
<td>Decision</td>
<td>Retain</td>
<td>Retain</td>
<td>Reject</td>
<td>Reject</td>
</tr>
</tbody>
</table>

The two hypotheses based on Experiment 1 were used to test category reliability and validity. No significant differences between the two groups were found for either of these hypothesis tests. Therefore, both null hypotheses were retained. The two hypotheses based on Experiment 2 were used to test coding reliability and validity. Significant differences were found for both of these tests. Therefore, both null hypotheses were rejected.

An analysis of those who did or did not complete Experiment 1 indicated differential participation by class enrollment. An analysis of those who did or did not complete Experiment 2 indicated differential participation by class enrollment and degree program.
CHAPTER V

DISCUSSION

Introduction

This chapter is used to present the final discussion of the results of the two experiments, along with some related issues. The five main topics covered here are: (1) an interpretation of the experimental results, (2) the limitations of the study, (3) benefits of the study, (4) suggestion for future research, and (5) a few closing comments.

Interpretation of the Results

Both the null hypotheses for Experiment 1 were retained. Thus, it is concluded the word count list and responses sorted by category identifiers and summaries did not help experimental participants create more reliable and valid category systems for the responses to the open-ended survey question used in the simulated evaluation effort.

On the other hand, both null hypotheses for Experiment 2 were rejected. Thus, it is also concluded the word count list and sorted responses did help experimental participants more reliably and validly code the open-ended responses into the final set of categories.

In terms of the general model for conducting an evaluation effort, this means the use of a word count list (key words out of
context) and responses sorted by category identifiers and summaries (information retrieval) were differentially effective. For Experiment 1, they did not help delineate higher quality category systems. However, for Experiment 2, they did help obtain higher quality codes for responses based on an established category system.

Limitations of the Study

Three considerations are presented in this section. First, limitations in the design of the study are discussed. Second, plausible explanations for retention of the null hypotheses of Experiment 1 are provided. Third, plausible explanations for differential participation in both experiments are identified.

Design Limitations

The ideal referent situation for both experiments would have been a real-world evaluation study that solicited responses to an open-ended survey question. Unfortunately, no such evaluation was available. The development of a contrived simulation activity using students enrolled in several College of Education courses was the best approximation available under the circumstances.

However, the demographic information on the study participants indicates these individuals had very little or no experience related to content analysis or other types of studies. As a result, the generalizability of the results do not relate directly to those people who typically perform these studies, that is, practicing researchers and evaluators.
Plausible Explanations for the Retention of the Null Hypotheses of Experiment 1

Both null hypotheses for Experiment 1 were retained while both null hypotheses for Experiment 2 were rejected. The exact reasons for these results can never be known, but it is assumed the different results are related to some differences between the two experiments. As a result, the identification of these differences can be used to help identify corresponding explanations for retention of the null hypotheses of Experiment 1. Two types of plausible explanations are discussed here: (1) the statistical power to detect true differences when they, in fact, exist, and (2) the characteristics of crucial elements of the experiments.

Statistical Power

As noted in Chapter 3, the desired sample size was established such that (a) the power of detecting a statistically significant difference between treatment groups when a difference, in fact, exists would be 0.60, and (b) the corresponding Type II error level would be 0.40. This level of power required at least 28 individuals to participate in each group (Cohen, 1977, p. 333).

For Experiment 1, the smaller of the two treatment groups finished with 32 members. Consequently, the actual power of detecting a true difference between treatment groups for this experiment was at least 0.65 (p. 333). For Experiment 2, the smaller group finished with 29 members. Thus, the actual power of detecting a true difference between treatment groups for Experiment 2 was at least 0.61
As a result, the statistical power for detecting true differences for Experiment 1 turned out to be higher than the power to detect true differences for Experiment 2.

Nevertheless, the probability of retaining the null hypotheses for Experiment 1 when differences between true scores do, in fact, exist was still 0.35. Even though this value is lower than originally planned, it is still a plausible explanation for retention of the null hypotheses of Experiment 1 at that level of probability.

**Characteristics of Crucial Elements**

The ability of a given treatment condition for an independent variable to change a participant's behavior, as measured through a dependent variable, is an indication of that treatment condition's potency. The purpose of conducting an experiment is to determine if one or more treatment conditions are sufficiently potent to differentially effect measures of one or more dependent variables obtained from different participant groups. By making features of the experimental procedures not related to the independent variable constant for all participants, and randomly assigning participants to different treatment groups; statistically significant differences between obtained scores on the dependent variables can reasonably be attributed to differences in treatment conditions—their differential potency.

Each experiment was designed to compare two different treatment conditions—performing a task with specialized microcomputer output vs. performing the task without the specialized output. The
experimental group worked under the first condition and the control group worked under the second condition. The results of the two experiments have already supported the decision the specialized outputs were not more potent for Experiment 1 but they were more potent for Experiment 2. Four possible sources for this difference in potency are discussed next. They are differences in (1) characteristics of the independent variables, (2) characteristics of the basic tasks, (3) characteristics of the participants, and (4) interactions between the three previous factors.

**Characteristics of the Independent Variables.** The independent variables for the two experiments were quite similar. In fact, the independent variable for Experiment 1 was a subset of the independent variable for Experiment 2. Thus, Experiment 2 was twice as long as Experiment 1. Furthermore, the additional portion of the Experiment 2 independent variable was directly analogous to the information retrieval procedures used in Experiment 1—sorting responses by category and including the appropriate category identifier and summary at the top of each group of similar responses. Consequently, the rather small differences in the independent variables for the two experiments are not considered to be a primary factor contributing to the different results. As a result, the likelihood of the Experiment 1 independent variable being the sole reason for retaining the null hypotheses, other than time, is also considered to be unlikely.

**Characteristics of the Basic Tasks.** The basic task for Experiment 1 was to create a new category system for a set of responses to an open-ended survey question. The participants were told in advance
all the responses were very negative in tone. They were also given a set of specifications for writing the category identifiers and summaries. The identifiers were to be short labels and the summaries were to be descriptive statements that clearly specified the negative nature of the responses in that category. The basic task for Experiment 2 was to code the set of responses in terms of an established set of categories for all participants. Participants only needed to assign one of the established category codes to each response.

Clearly, these two tasks are quite different. Not only are they substantively different, but it is also likely the category development task was more difficult than the response coding task. Two basic reasons for this greater difficulty seem likely.

First, it might be the result of the need to use higher level intellectual abilities and skills to develop a category system versus only coding responses based on an established system. In terms of Bloom's Taxonomy of Educational Objectives (Bloom et al., 1956), the task of developing a category system involves a complex process that requires a number of intellectual abilities and skills at the three highest levels of the taxonomy—4.00 analysis, 5.00 synthesis, and 6.00 evaluation (pp. 144-200). The task of coding responses based on an established category system requires abilities and skills at a lower level—3.00 application (pp. 120-143). Therefore, the need to use higher level skills is a plausible explanation for retention of the null hypotheses of Experiment 1.

Second, the specific nature of this particular category system development task also could have contributed to its difficulty. A
large number of uses of content analysis were presented in Chapter 2. In addition, several types of coding units were described. The purpose of this particular study was to describe the attitudes of a group of teachers about a highly controversial accountability system used to evaluate their performance. The analysis also used rather complex natural language coding units ranging in length from a sentence to a short paragraph. The categories the participants had to create were essentially themes or assertions about some aspect of the accountability system. Thus, it is also plausible this aspect of the Experiment 1 tasks contributed to retention of the null hypotheses.

**Characteristics of the Participants.** If the characteristics of the participants who completed Experiment 1 were somehow different than the characteristics of the participants who completed Experiment 2, these characteristics might be related to retention of the null hypotheses of Experiment 1. The only way this could happen for this study is if the drop-out pattern for Experiment 1 were different than the drop-out pattern for Experiment 2. These patterns were analyzed and the complete results were presented in Chapter 4.

It turns out the drop-out patterns for Experiment 1 and Experiment 2 were different. Those who did or did not complete Experiment 1 differed in which class they attended. No other differences were found between the two groups. For Experiment 2, the same difference was found plus one more. A disproportionately high number of graduate students completed Experiment 2 compared to the proportion of undergraduate students who completed Experiment 2. As a result, certain participant characteristics might have also contributed to
retention of the null hypotheses of Experiment 1. Specifically, a relatively high proportion of undergraduate students who completed Experiment 1 might have contributed to the nonsignificant results.

Interactions Between the Characteristics of the Independent Variables, Basic Tasks, and Participants. The most plausible explanation for the nonsignificant results of Experiment 1 is some form of interaction between the three factors discussed above. For whatever reasons, the specialized microcomputer output used in Experiment 1 to help develop a category system by the participants involved was not sufficiently potent to produce reliability or validity scores significantly higher than those obtained by the control group. Suggestions for further research designed to help identify the role of these factors in improving the quality of content analyses used in evaluation efforts are presented in a later section.

Plausible Explanations for Differential Participation

The composition of the groups who did or did not complete each experiment were found to be different on some characteristics. For Experiment 1, differential participation was indicated by class enrollment. Two classes had a combined participation rate of about 90%, while the four remaining classes had a combined participation rate of 50%. This difference was also found for Experiment 2. For that experiment, the first two classes had a combined participation rate of about 86%, and the four remaining classes had a combined participation rate of just over 35%. The groups of those who did or did not complete Experiment 2 also differed on a second characteristic.
About 62% of all graduate students completed Experiment 2, while less than 40% of all undergraduate students completed it.

This differential participation limits the generalizability of the results in unknown ways. However, controlled speculation about the reasons for the above circumstances can provide some clues to how they can be avoided or explicitly studied in the future. The remainder of this section covers differential participation in terms of two characteristics of people in the study group: (1) class enrollment and (2) degree program.

**Class Enrollment**

A plausible explanation for differential participation related to class enrollment is readily available. The students enrolled in the higher participation classes were given a different incentive by their instructors than the students enrolled in the four lower participation classes. The students enrolled in the higher participation classes were allowed to replace one regular assignment (a paper in each class) with participation in all four tasks of the study. Also, they automatically received an "A" for that assignment. The students enrolled in the lower participation classes were given what was then considered to be a weaker incentive. They would be given the higher of two possible final grades for the class in "borderline" cases. The differential participation results indicate this incentive was, in fact, much weaker.

This difference in incentives appears to be the prime reason for the differential participation by class membership for the two
experiments. The apparent strengths of the incentives (86% to 90% participation for assignment substitution vs. 35% to 50% participation for "benefit of the doubt" grading) should also be heeded by designers of future research efforts that depend on the voluntary participation of university students. The lesson appears to be this. It is better to have a relatively few participants with strong incentives to complete the study than a higher number of participants with weak incentives to complete the study.

**Degree Program**

Unlike class enrollment, differential participation in Experiment 2 by degree program is not so easily explained. As noted earlier, about 62% of the graduate students and less than 40% of the undergraduate students completed Experiment 2. However, it turns out only one person from the higher participation classes who completed Experiment 1 (less than 4% of that subgroup) did not complete Experiment 2. On the other hand, 14 people from the lower participation classes who completed Experiment 1 (over 29% of that subgroup) did not complete Experiment 2. Thus, the people who did complete Experiment 1 but did not complete Experiment 2 were, for the most part, enrolled in the classes that offered the weaker incentive.

As a result, differential participation in Experiment 2 by degree program might somehow be related to the strength of the incentive to participate. For example, graduate students might be more likely to complete a study with weak incentives to participate than are undergraduate students, given the same weak incentives. Possible
reasons for this might be related to typical differences in the sources and strengths of incentives for graduate and undergraduate students to complete long, difficult tasks. Graduate students might be accustomed to having few or weak external sources of incentives, forcing them to be "self-motivated," while undergraduate students might be more accustomed to being offered stronger incentives by other people before they perform such tasks. When those incentives are not offered, perhaps they are less likely to perform. For whatever reasons, graduate students turned out to be more durable than undergraduate students in this study.

Benefits of the Study

This study can provide useful information to three groups of people. These groups include: (1) practicing evaluators interested in conducting content analyses of responses to open-ended survey questions, (2) researchers interested in conducting experiments that require reliability and validity measures directly related to unique simulation problems, and (3) theoreticians interested in studying the relationships between evaluation, content analysis, and microcomputers.

The two experiments address the problem of analyzing a set of responses to an open-ended survey question used in a simulated evaluation effort. Experiment 1 focused on developing a new category system, and Experiment 2 focused on coding responses into an established set of categories. Through these experiments, the effects of specialized microcomputer output—using a word count list plus
responses sorted by category and headed with the applicable identifier and summary—on the reliability and validity of the resultant category system or codes were tested. No effects were found for the category experiment. However, the specialized microcomputer output was found to help the experimental participants produce more reliable and valid codes for the responses. Thus, practicing evaluators might benefit in two ways from using such specialized output to code responses to open-ended survey questions. First, studies that would have been conducted even without the availability of such output might be conducted more reliably with more meaningful results. Second, new studies might be conducted that otherwise would have been considered too difficult to conduct by non-computerized methods.

Researchers who attempt to study some aspect of a real-world problem through a simulation activity are more likely to represent the holistic nature of that problem than if they had used a highly controlled laboratory experiment. The simulation's uniqueness also has the disadvantage of precluding highly standardized measures of dependent variables from being available. The lack of pre-existing reliability and validity measures for this study was addressed through the use of two panels of education and evaluation experts. These panels generated the necessary criteria and scored the experimental data in accordance with those criteria. The methods used in these activities are general enough that researchers conducting similar studies can adapt them to their own situations.

Those who are interested in theoretical considerations might gain a better understanding of the relationships between evaluation,
content analysis, and microcomputers. They might also be encouraged to pursue related lines of research on how microcomputers can be used to enhance the understanding of and practice in each of these fields.

Suggestions for Future Research

Three elements crucial to the success of Experiment 1 were identified during the discussion of why its corresponding null hypotheses were retained. These elements were characteristics of the: (1) independent variable, (2) basic category task, and (3) participants. Future studies that focus on these elements as independent variables might help determine how they are related to producing category systems and codes for responses to open-ended survey questions that are both reliable and valid. Possible levels of these characteristics are discussed next.

Each independent variable for this study had two basic components: (1) a word count list and (2) responses sorted by categories and headed by their corresponding identifiers and summaries. This variable—microcomputer output—could be divided into four levels for future studies: (1) no specialized output, (2) word list only, (3) sorted responses only, and (4) both word list and sorted responses.

The basic task for Experiment 1 was to develop a category system for a collection of statements made by a group of teachers about their school's accountability system. The coding units tended to be rather long—one to a few sentences—and the categories were expected to reflect a general theme about the accountability system. Each theme was also required to take the form of a negative assertion. A
shorter and simpler type of coding unit often found in responses to open-ended survey questions is the word or small group of words. For example, the question, "What is your primary source of information about the school system?" will tend to promote very short responses. The categories for responses to such a question are usually very similar to the actual responses. As such, they do not represent themes or assertions but simply designations of entities. Because of this difference, these categories might also be easier to develop than categories based on assertions. Thus, a category development task variable could have two levels: (1) designations and (2) assertions.

Finally, participants of the study were undergraduate and graduate students enrolled in several College of Education classes. It was not possible to compare how these two groups performed in this study, but graduate students did have more success completing it. In addition, all these people turned out to have little or no research or evaluation related experience. As a result, it is difficult to speculate exactly how well practicing evaluators would have performed the simulation tasks. In any event, three levels of a participant variable might help sort out any important differences between: (1) undergraduate students, (2) graduate students, and (3) practicing evaluators.

All three potential independent variables could be combined into a single study using a three-factor design with four, two, and three levels, respectively. For consistency, the dependent variables could be comparable to those used for this study. However, according to
Cohen (1977), such a 4 X 2 X 3 factorial design with alpha set at 0.05, power set at 0.80, and a medium effect size would require 32 participants per cell (p. 321) or 768 participants overall. This breaks down to 256 people from each participant group. Based on the problems this researcher had getting local participants to complete this study, the prospects of getting 256 practicing evaluators to finish the study—let alone finding them—seem laughable.

In a word, the solution to this dilemma seems to be this—compromise. Of all the possible comparisons, a relatively few of them seem to be of more interest than the others. For the microcomputer output factor, the comparison could be between using the word list alone vs. using the sorted responses alone. For the category task factor, only one comparison is possible—designations vs. assertions. For the participant factor, two comparisons seem interesting: (1) undergraduate vs. graduate students and (2) any students vs. practicing evaluators. Three designs could be used to study these comparisons. However, the exact nature of these studies would depend on the actual contexts in which they were conducted and the resources available.

The first study could use a two-factor design: (1) designations vs. assertions by (2) undergraduate vs. graduate students. The computer output could include both levels for this study. Using the above values for alpha, power, and effect size, this study would require only 248 students (Cohen, 1977, p. 312). In addition, it would require at least two different open-ended questions be used, one like the question in this section and one like the question used
in the simulation. This design also allows for testing interaction effects between the level of task and level of participant.

The second study could also use a two-factor design: (1) designations vs. assertions and (2) list only vs. sorted responses only. The participants could be a mixture of undergraduate and graduate students. Again, 248 students would be required. Repeating the task factor would also allow a different set of open-ended questions to be used. This allows for testing interaction effects between the level of task and level of microcomputer output.

The third study could use a one-factor design at three levels: (1) undergraduate students vs. (2) graduate students vs. (3) practicing evaluators. All participants could receive both levels of specialized microcomputer output and the more difficult of the two levels of tasks (as established by the above experiments) could be used. For this study, 156 participants would be required (Cohen, 1977, p. 314), but only 52 from each participant group. This is the lowest required number of practicing evaluators for any of the possible designs related to at least one of the three factors and using the above specifications for alpha, power, and effect size. Therefore, this design holds the best chance for learning how well practicing evaluators can use specialized microcomputer output to develop category systems and code responses to open-end survey questions.

Closing Comments

The purpose of this study was to advance the body of knowledge about how evaluation practitioners can use microcomputer programs to
improve the reliability and validity of content analyses of responses to open-ended survey questions used in evaluation efforts. This purpose was to be achieved by identifying key relationships between evaluation, content analysis, and microcomputers; conducting two experiments on category system development or response coding; and making recommendations for practice and further study. This researcher believes the purpose and objectives of this study have been reached, although it was a long and difficult path.

From this vantage point, it is easy to say certain things could have been performed better or differently. However, even without the benefit of such hindsight, the study was planned and conducted so that, at every stage of its development, it could always make the most of its potential.

Because of the major role content analysis played in this study, it is only fitting this chapter closes with an adaptation of the final comment made by Berelson (1952) in his classic work on the subject. Even if nothing else was accomplished by this study, perhaps it has shown evaluation and content analysis have no magic qualities. "You rarely get out of [them] more than you put in, and sometimes you get less. In the last analysis, there is no substitute for a good idea" (p. 198).
Appendix A

Development of the Response Pool
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 created during the activity provided the basis for judging the validity of the participants' content analyses of the responses. The response pool development tasks were: (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 count list of all 100 responses.

Select Source of Information

The first task was to select a general source of information from which a response pool could be developed. An ideal source of
responses would be a non-trivial, real-world evaluation report that contained at least one open-ended question and a large, diverse, unedited set of responses to such questions.

An evaluation report that meets these criteria is described by Patton (1980, pp. 23-30). The evaluation report included a summary of the findings from a mail survey given to the teachers of a public school district in the Midwest that used a controversial accountability system. The survey information was intended to help evaluate the accountability system from the perspective of the teachers. The questionnaire included a number of forced-choice questions and two open-ended questions. Three hundred seventy-three teachers (70 percent of those who responded to the questionnaire) responded to at least one of the open-ended questions. All of the comments written by the teachers were typed verbatim and included in the report, filling 101 single-spaced pages. The researcher obtained a copy of this evaluation report (Patton, French, & Perrone, 1976) and used it as the basis for developing the response pool.

Select Open-Ended Question

Next, one open-ended question was written for use in the simulation. The two questions used in the actual evaluation study and the responses elicited by them were sufficiently comparable to allow for one hybrid question to be used. This allowed all responses from the teachers to be considered when the simulation responses were developed. The two open-ended questions used in the study and the one open-ended question used in the simulation are presented in Table 1.
Table 1
Open-Ended Questions Used by Patton et al. (1976) and Open-Ended Question Used in the Simulation

<table>
<thead>
<tr>
<th>Item</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>II. c. (p. 7)</td>
<td>Please use this space to make any further comments or recommendations concerning any components of the accountability system.</td>
</tr>
<tr>
<td>IX. (p. 15)</td>
<td>Comments—Finally, we'd like you to use this space to add any additional comments you'd like to make about any part of the [Hometown] accountability system.</td>
</tr>
<tr>
<td>Simulation</td>
<td>Please give us any comments or recommendations you would like to make about any part of the Hometown Public Schools accountability system.</td>
</tr>
</tbody>
</table>

Determine Basic Classification Framework

Because the emphasis of this study was on using output from commercially available microcomputer programs to improve the reliability and validity of content analysis, the general framework used for developing a category system in the simulation was intentionally kept simple. A highly complex framework would tend to draw attention away from the intended emphasis, especially if the participants found the framework too difficult to use. Osgood, Suci, and Tannenbaum (1957) found that people often make distinctions between objects along three dimensions: (a) evaluation, such as, good-bad or positive-negative; (b) potency, such as, hard-soft or strong-weak; and (c) activity, such as, active-passive or fast-slow (pp. 62-63). For simplicity, the dimension most relevant to an evaluation study, evaluation, was selected to be used along with an attitude object.
dimension with components like teacher turnover or testing program. These two dimensions would then constitute a general framework for the panel and participants to create a category system.

Before the researcher read the set of responses in the actual evaluation report, the category system was to have the following structure for each category: [Evaluative] statement about [Object]; where [Evaluative] would be replaced with "Negative," "Neutral," or "Positive;" and [Object] would be replaced with a descriptive identifier for some aspect of the accountability system. It turns out that virtually every comment made by the 373 teachers about the accountability system was negative. This made the intended classification system inappropriate because one dimension would have no variability.

Instead, the general framework was modified so that each category was first given to represent a group of negative comments about the accountability system. Each category was then expected to have two parts: (1) an identifier—a brief, descriptive title for the category that indicates its negative nature and (2) a summary—an operational definition of the category complete enough to allow coders to decide whether a particular response should or should not be included in that category. It was also decided at that time to use 10 categories in the framework in order for the simulation problem to be challenging yet manageable for the participants. Because 10 categories and the related responses were eventually determined to be too many, the number of categories and responses was later cut in half. A graphic representation of the general framework for developing the category system is presented in Figure 1.
At this point, the potential responses for the panel to review were prepared. The goal of the panel was to select and code 200 responses that would fit into 10 categories they developed to comply with the general framework just described. One hundred responses from five categories were later selected as the final response pool. The researcher facilitated this process by preparing an oversized pool of specially designed responses for the panel to review. These responses were prepared through an iterative process as follows: (a) several category identifiers and summaries, based on the Patton et al. (1976) evaluation report, were drafted on numbered index cards; (b) responses or portions of responses that fit a category were
selected and modified, if necessary; (c) the respondent identification number was written on the card; (d) a data base of responses was constructed on a DECmate II list processing program with each record identifying the category number, respondent number, and text of each response; (e) the data base was occasionally sorted by category and printed for review and modification; (f) as a new category became apparent a card was prepared for it; (g) new categories and responses were added until 15 categories contained about 300 responses with about 20 responses per category; (h) 10 categories were then selected from the 15 and new responses were added until there were about 200 responses with roughly 20 responses per category.

The following criteria were used for writing and selecting the categories: (a) The categories should convey a negative evaluation about some aspect of the accountability system. Because neutral or positive responses were, by and large, not present in the set of responses, one content dimension and one evaluative dimension for each category was not needed. Both dimensions could be incorporated into a single, negatively weighted category. (b) The categories should be at roughly comparable levels of detail and fairly specific. This criterion should help raters decide into what category a response should be coded. (c) The categories should be exhaustive. All responses should fit somewhere. If this were not the case, a "garbage" category like "other" would be needed. This would not satisfy the other criteria. (d) The categories should be mutually exclusive. One best category for each response should exist. One category per response greatly simplifies analysis and interpretation.
(e) The categories should require more than a key word or phrase in most responses in order for a response to be properly coded. This criterion was intended to raise the difficulty of the tasks to a sufficiently high level to allow for differential performance to be demonstrated between the experimental and control groups. This was accomplished by having key words such as "teachers" be relevant to more than one category or by avoiding category definitions that could be operationally simplified to one key word.

A complementary set of criteria was used for writing and selecting the responses. The responses should: (a) convey a negative evaluation about some aspect of the accountability system, (b) have a clear best category into which it should be coded, (c) collectively not be codable by key words or phrases alone, and (d) be no more than a few sentences long. These criteria represent the iterative process necessary to produce a coherent yet realistic set of categories and responses at a usable level of difficulty for the experiments to follow.

Recruit a Panel

A group of four faculty members from Western Michigan University's College of Education was recruited to serve on the Response Panel. These faculty members included the Chair of the Departments of Educational Leadership, and Counseling and Personnel; one professor of Educational Leadership; one professor of Educational and Professional Development; and one associate professor of Educational and Professional Development. All panelists have had extensive
experience in public school settings and various types of educational research and evaluation activities.

**Have Panelists Independently Create a Category System**

The panelists received written instructions about how to proceed. They could also ask the researcher questions as needed. The panelists were given two weeks to create 10 categories and code 204 responses. Materials were returned to the researcher within 10 days.

The simulation materials used by the Response Panel were analogous to those used during the first task of the experiments. The main difference was Response Panel members were asked to create 10 categories and code about 200 responses while the participants were asked to create five categories and code 50 responses in Task 1.

The Response Panel received the following materials along with their written instruction: (a) Read Me First—an overview of the simulation and the simulation instructions, (b) Draft Introduction—a description of the background of the problem, (c) Practice Exercise—an in-class warm-up exercise, (d) Practice Exercise Answer Sheet—sample answers to the exercise, (e) Sample of Responses—204 responses to be coded, (f) Category Development Worksheets—worksheets for writing category identifiers and summaries, and (g) all 204 responses repeated on individual cards with 10 envelopes—a sorting aid not provided to the participants. Items (a) through (d) are presented in Appendix D. Items (e) and (f) are presented in Appendix B.
Process the Panelists' Work

The researcher processed the categories and response codes of the panelists and researcher in order to facilitate the group activity. This processing consisted of: (a) entering the category identifiers and summaries into a data base on a DECMate II; (b) mapping each of the panelists' categories onto the researcher's categories by labeling the researcher's categories "A" through "J" and then assigning each panelist's numeric category to the alphabetic label with the closest definition; (c) translating each panelist's numeric code for each response to the corresponding alphabetic code for the response; (d) entering the response codes onto another data base on the DECMate II; (e) sorting category identifiers and summaries by the alphabetic codes and printing them as the Response Panel Category Summary; (f) sorting the response codes by response identification number and by the code most used for each response; (g) printing both lists without the text of the responses as the Response Panel Classifications of Responses; (h) printing the alphabetic codes sorted by response identification number with the text of the responses; (i) constructing and printing a frequency coincidence matrix, and a related measure of agreement by categories, Krippendorff's alpha (Krippendorff, 1980, pp. 140-146), and, finally, (j) constructing and printing a percentage coincidence matrix by category, along with a corresponding Venn diagram. All printed materials were handed out to the panelists at the start of the group meeting.
Define Categories and Assign Responses

The four faculty members and the researcher met with the intention to finalize the category system and set of responses. However, because of the circumstances described in the next section, one more selection activity was needed. This group of five people constituted the Response Panel as far as decisions at the meeting were concerned. The first task was to finalize the category identifiers and summaries. Panelists discussed each category and the researcher wrote each identifier and summary pair on separate sheets of posted newsprint as each decision was made. The next task was to finalize the code and wording for 200 responses and discard four responses. Four out of five panelist had to agree on the disposition of a response before it was finalized. Responses with less than 80 percent agreement were highlighted on the appropriate handouts—26 out of 204 responses. After exactly 200 responses were approved the meeting was adjourned.

Select the Final Categories and Responses

The day after the panel meeting, the researcher cut the response pool down to 100 responses and five categories because of concern the tasks planned for the experiments would be too complex and time consuming for the participants. This decision was made because (a) Response Panel members expressed concern that the response pool probably was too large, (b) the first pilot study instructor dropped out of the study after seeing the size of the response pool, and (c) the day after the Response Panel meeting, the second pilot study
instructor, after seeing the size of the response pool, withheld consent to participate in the experiment until an indefinite later date. The criteria for selecting the final categories and responses were the same as stated above with the added criterion that the first task undertaken by participants in the experiments should take no more than one hour to complete. Time reports from the pilot study indicated five categories and 100 responses met this criterion.

**Divide the Response Pool Into Two Groups**

Responses were randomly assigned to two groups of 50 responses each. One group of responses was arbitrarily designated to be used with the first experiment. Both groups of responses were used in the second experiment.

**Produce a Word Count List**

The Word Count List was created from a microcomputer text file that contained the final 100 responses. This text file was processed using the "WORDFREQ" option of The Word Plus (Holder, 1982, p. 38) spelling checker program running on an Osborne 1 microcomputer. The resultant list was ordered in descending frequency of occurrence of each unique word. This list file was transferred to a DECmate II microcomputer where all words occurring only once were deleted, an appropriate heading was added, and the list was arranged in columns. The resultant document is presented as Exhibit 1.
Here is a word count list I asked my secretary to put together from all of the responses to the open-ended question. He created it with one of the options on our spelling checker program on the word processor. The list contains all of the words that appeared more than once, sorted in order by frequency of occurrence. Maybe these lists will give you a few leads to follow when you start to develop the five categories of responses.

<table>
<thead>
<tr>
<th>#</th>
<th>WORD</th>
<th>#</th>
<th>WORD</th>
<th>#</th>
<th>WORD</th>
</tr>
</thead>
<tbody>
<tr>
<td>108</td>
<td>THE</td>
<td>9</td>
<td>WOULD</td>
<td>5</td>
<td>SO</td>
</tr>
<tr>
<td>72</td>
<td>TO</td>
<td>9</td>
<td>WHICH</td>
<td>5</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>59</td>
<td>AND</td>
<td>9</td>
<td>THEY</td>
<td>5</td>
<td>VARIABLES</td>
</tr>
<tr>
<td>55</td>
<td>IN</td>
<td>9</td>
<td>Peer</td>
<td>5</td>
<td>MUST</td>
</tr>
<tr>
<td>53</td>
<td>A</td>
<td>9</td>
<td>THEIR</td>
<td>5</td>
<td>TESTS</td>
</tr>
<tr>
<td>47</td>
<td>OF</td>
<td>9</td>
<td>OUR</td>
<td>5</td>
<td>ETC</td>
</tr>
<tr>
<td>44</td>
<td>IS</td>
<td>9</td>
<td>OTHER</td>
<td>5</td>
<td>FAR</td>
</tr>
<tr>
<td>40</td>
<td>I</td>
<td>9</td>
<td>TOO</td>
<td>5</td>
<td>USED</td>
</tr>
<tr>
<td>36</td>
<td>ARE</td>
<td>8</td>
<td>BY</td>
<td>5</td>
<td>WORK</td>
</tr>
<tr>
<td>34</td>
<td>ACCOUNTABILITY</td>
<td>8</td>
<td>BEEN</td>
<td>5</td>
<td>AM</td>
</tr>
<tr>
<td>33</td>
<td>TEACHERS</td>
<td>8</td>
<td>WAS</td>
<td>5</td>
<td>S</td>
</tr>
<tr>
<td>32</td>
<td>NOT</td>
<td>8</td>
<td>WORKING</td>
<td>5</td>
<td>LEARNING</td>
</tr>
<tr>
<td>31</td>
<td>BE</td>
<td>8</td>
<td>ACCOUNTABLE</td>
<td>5</td>
<td>WANTS</td>
</tr>
<tr>
<td>28</td>
<td>AS</td>
<td>8</td>
<td>STUDENTS</td>
<td>5</td>
<td>ANY</td>
</tr>
<tr>
<td>28</td>
<td>IT</td>
<td>8</td>
<td>ADMINISTRATION</td>
<td>5</td>
<td>SEEMS</td>
</tr>
<tr>
<td>28</td>
<td>THAT</td>
<td>8</td>
<td>IF</td>
<td>4</td>
<td>RATE</td>
</tr>
<tr>
<td>26</td>
<td>SYSTEM</td>
<td>7</td>
<td>ONE</td>
<td>4</td>
<td>RATING</td>
</tr>
<tr>
<td>23</td>
<td>FOR</td>
<td>7</td>
<td>SOME</td>
<td>4</td>
<td>HOWEVER</td>
</tr>
<tr>
<td>19</td>
<td>WE</td>
<td>7</td>
<td>RATINGS</td>
<td>4</td>
<td>SHOULD</td>
</tr>
<tr>
<td>18</td>
<td>WITH</td>
<td>7</td>
<td>IDEA</td>
<td>4</td>
<td>BELIEVE</td>
</tr>
<tr>
<td>18</td>
<td>BUT</td>
<td>7</td>
<td>PEOPLE</td>
<td>4</td>
<td>SCORES</td>
</tr>
<tr>
<td>18</td>
<td>HAVE</td>
<td>7</td>
<td>WHEN</td>
<td>4</td>
<td>AT</td>
</tr>
<tr>
<td>16</td>
<td>GOOD</td>
<td>7</td>
<td>THIS</td>
<td>4</td>
<td>WILL</td>
</tr>
<tr>
<td>16</td>
<td>ON</td>
<td>6</td>
<td>MANNER</td>
<td>4</td>
<td>INDIVIDUAL</td>
</tr>
<tr>
<td>13</td>
<td>WHO</td>
<td>6</td>
<td>OR</td>
<td>4</td>
<td>USEFUL</td>
</tr>
<tr>
<td>13</td>
<td>HOMETOWN</td>
<td>6</td>
<td>MAKE</td>
<td>4</td>
<td>THOSE</td>
</tr>
<tr>
<td>12</td>
<td>HAS</td>
<td>6</td>
<td>DOES</td>
<td>4</td>
<td>CHILDREN</td>
</tr>
<tr>
<td>12</td>
<td>CAN</td>
<td>6</td>
<td>HE</td>
<td>4</td>
<td>TESTING</td>
</tr>
<tr>
<td>12</td>
<td>DO</td>
<td>6</td>
<td>BECAUSE</td>
<td>4</td>
<td>BAD</td>
</tr>
<tr>
<td>11</td>
<td>NO</td>
<td>6</td>
<td>SCHOOL</td>
<td>4</td>
<td>HIGH</td>
</tr>
<tr>
<td>11</td>
<td>THERE</td>
<td>6</td>
<td>OUT</td>
<td>4</td>
<td>JUST</td>
</tr>
<tr>
<td>10</td>
<td>ALL</td>
<td>6</td>
<td>WAY</td>
<td>4</td>
<td>EDUCATIONAL</td>
</tr>
<tr>
<td>10</td>
<td>AN</td>
<td>5</td>
<td>YOU</td>
<td>4</td>
<td>SEE</td>
</tr>
<tr>
<td>10</td>
<td>HUMAN</td>
<td>5</td>
<td>EXCELLENT</td>
<td>4</td>
<td>AMONG</td>
</tr>
<tr>
<td>10</td>
<td>TEACHER</td>
<td>5</td>
<td>THAN</td>
<td>4</td>
<td>EXPECTED</td>
</tr>
<tr>
<td>10</td>
<td>FEEL</td>
<td>5</td>
<td>LIKE</td>
<td>4</td>
<td>HAT</td>
</tr>
<tr>
<td>10</td>
<td>MANY</td>
<td>5</td>
<td>PROGRAM</td>
<td>4</td>
<td>US</td>
</tr>
<tr>
<td>10</td>
<td>EACH</td>
<td>5</td>
<td>WERE</td>
<td>4</td>
<td>FROM</td>
</tr>
<tr>
<td>#</td>
<td>WORD</td>
<td>#</td>
<td>WORD</td>
<td>#</td>
<td>WORD</td>
</tr>
<tr>
<td>---</td>
<td>--------------</td>
<td>---</td>
<td>--------------</td>
<td>---</td>
<td>--------------</td>
</tr>
<tr>
<td>4</td>
<td>ANOTHER</td>
<td>3</td>
<td>BECOME</td>
<td>2</td>
<td>UPON</td>
</tr>
<tr>
<td>4</td>
<td>BEST</td>
<td>3</td>
<td>LEVEL</td>
<td>2</td>
<td>DIDN'T</td>
</tr>
<tr>
<td>4</td>
<td>EVEN</td>
<td>3</td>
<td>GOAT</td>
<td>2</td>
<td>ITS</td>
</tr>
<tr>
<td>4</td>
<td>TEACHING</td>
<td>3</td>
<td>MOST</td>
<td>2</td>
<td>COMPONENTS</td>
</tr>
<tr>
<td>4</td>
<td>INTO</td>
<td>3</td>
<td>I'M</td>
<td>2</td>
<td>WHAT</td>
</tr>
<tr>
<td>4</td>
<td>STUDENT</td>
<td>3</td>
<td>RESULT</td>
<td>2</td>
<td>HELD</td>
</tr>
<tr>
<td>4</td>
<td>TAKE</td>
<td>3</td>
<td>LOWER</td>
<td>2</td>
<td>FIGHTING</td>
</tr>
<tr>
<td>3</td>
<td>TOGETHER</td>
<td>3</td>
<td>VERY</td>
<td>2</td>
<td>GONE</td>
</tr>
<tr>
<td>3</td>
<td>SHARING</td>
<td>3</td>
<td>IDEAS</td>
<td>2</td>
<td>DECENCY</td>
</tr>
<tr>
<td>3</td>
<td>HIS</td>
<td>3</td>
<td>THEM</td>
<td>2</td>
<td>STANDARDIZED</td>
</tr>
<tr>
<td>3</td>
<td>IT'S</td>
<td>2</td>
<td>CHECK</td>
<td>2</td>
<td>DEMEANING</td>
</tr>
<tr>
<td>3</td>
<td>OBJECTIVES</td>
<td>2</td>
<td>ABOUT</td>
<td>2</td>
<td>PROBLEM</td>
</tr>
<tr>
<td>3</td>
<td>RATED</td>
<td>2</td>
<td>SELF</td>
<td>2</td>
<td>DURING</td>
</tr>
<tr>
<td>3</td>
<td>NONE</td>
<td>2</td>
<td>HIM</td>
<td>2</td>
<td>AFRAID</td>
</tr>
<tr>
<td>3</td>
<td>UP</td>
<td>2</td>
<td>NEED</td>
<td>2</td>
<td>PROGRESS</td>
</tr>
<tr>
<td>3</td>
<td>THINK</td>
<td>2</td>
<td>PRODUCT</td>
<td>2</td>
<td>DOESN'T</td>
</tr>
<tr>
<td>3</td>
<td>SHOW</td>
<td>2</td>
<td>ACHIEVE</td>
<td>2</td>
<td>RESENT</td>
</tr>
<tr>
<td>3</td>
<td>ONLY</td>
<td>2</td>
<td>SAME</td>
<td>2</td>
<td>JOB</td>
</tr>
<tr>
<td>3</td>
<td>DONE</td>
<td>2</td>
<td>EVALUATIONS</td>
<td>2</td>
<td>JOKE</td>
</tr>
<tr>
<td>3</td>
<td>CONTROL</td>
<td>2</td>
<td>CONSIDERATION</td>
<td></td>
<td>INCREASING</td>
</tr>
<tr>
<td>3</td>
<td>PRINCIPAL</td>
<td>2</td>
<td>B</td>
<td>2</td>
<td>CLASSES</td>
</tr>
<tr>
<td>3</td>
<td>PRESSURE</td>
<td>2</td>
<td>IMPLEMENTED</td>
<td>2</td>
<td>NEVER</td>
</tr>
<tr>
<td>3</td>
<td>THINGS</td>
<td>2</td>
<td>BUSINESS</td>
<td>2</td>
<td>JUDGING</td>
</tr>
<tr>
<td>3</td>
<td>CONCERNED</td>
<td>2</td>
<td>ATTITUDES</td>
<td>2</td>
<td>TOOL</td>
</tr>
<tr>
<td>3</td>
<td>DEALING</td>
<td>2</td>
<td>BACKGROUNDS</td>
<td>2</td>
<td>EVALUATION</td>
</tr>
<tr>
<td>3</td>
<td>FORCE</td>
<td>2</td>
<td>ORDER</td>
<td>2</td>
<td>LOST</td>
</tr>
<tr>
<td>3</td>
<td>CAUSE</td>
<td>2</td>
<td>MIGHT</td>
<td>2</td>
<td>ROOM</td>
</tr>
<tr>
<td>3</td>
<td>AGAINST</td>
<td>2</td>
<td>SIMPLY</td>
<td>2</td>
<td>WELL</td>
</tr>
<tr>
<td>3</td>
<td>ADMINISTRATORS</td>
<td>2</td>
<td>SLOW</td>
<td>2</td>
<td>TIME</td>
</tr>
<tr>
<td>3</td>
<td>CHILD</td>
<td>2</td>
<td>POSITIVE</td>
<td>2</td>
<td>YEAR'S</td>
</tr>
<tr>
<td>3</td>
<td>BETWEEN</td>
<td>2</td>
<td>HUMANITY</td>
<td>2</td>
<td>CARE</td>
</tr>
<tr>
<td>3</td>
<td>GROWTH</td>
<td>2</td>
<td>GET</td>
<td>2</td>
<td>CLASS</td>
</tr>
<tr>
<td>3</td>
<td>BEINGS</td>
<td>2</td>
<td>POSSIBLE</td>
<td>2</td>
<td>WHOLE</td>
</tr>
<tr>
<td>3</td>
<td>MORE</td>
<td>2</td>
<td>RESPECT</td>
<td>2</td>
<td>UNDER</td>
</tr>
<tr>
<td>3</td>
<td>COULD</td>
<td>2</td>
<td>DON'T</td>
<td>2</td>
<td>HOME</td>
</tr>
<tr>
<td>3</td>
<td>ME</td>
<td>2</td>
<td>TRYING</td>
<td>2</td>
<td>SINCE</td>
</tr>
<tr>
<td>3</td>
<td>MY</td>
<td>2</td>
<td>EVERYONE</td>
<td>2</td>
<td>PLAY</td>
</tr>
<tr>
<td>3</td>
<td>EVER</td>
<td>2</td>
<td>PROBLEMS</td>
<td>2</td>
<td>LOW</td>
</tr>
<tr>
<td>3</td>
<td>SUPERINTENDENT</td>
<td>2</td>
<td>MANAGEMENT</td>
<td>2</td>
<td>MERIT</td>
</tr>
<tr>
<td>3</td>
<td>WRONG</td>
<td>2</td>
<td>TRY</td>
<td>2</td>
<td>ECONOMIC</td>
</tr>
<tr>
<td>3</td>
<td>COMPETITION</td>
<td>2</td>
<td>SHORT</td>
<td>2</td>
<td>PAY</td>
</tr>
<tr>
<td>3</td>
<td>RATHER</td>
<td>2</td>
<td>THING</td>
<td>2</td>
<td>BETTER</td>
</tr>
<tr>
<td>3</td>
<td>STAFF</td>
<td>2</td>
<td>VALUE</td>
<td>2</td>
<td>SHARE</td>
</tr>
<tr>
<td>3</td>
<td>CANNOT</td>
<td>2</td>
<td>IMPORTANT</td>
<td>2</td>
<td>THEMSELVES</td>
</tr>
<tr>
<td>3</td>
<td>GROUP</td>
<td>2</td>
<td>SITUATION</td>
<td>2</td>
<td>PART</td>
</tr>
<tr>
<td>3</td>
<td>HERE</td>
<td>2</td>
<td>MATERIALS</td>
<td>2</td>
<td>MEETING</td>
</tr>
<tr>
<td>3</td>
<td>LITTLE</td>
<td>2</td>
<td>EXPERIENCE</td>
<td>2</td>
<td>RESPONSIBLE</td>
</tr>
<tr>
<td>3</td>
<td>MODEL</td>
<td>2</td>
<td>MAY</td>
<td>2</td>
<td>OTHERS</td>
</tr>
<tr>
<td>3</td>
<td>USE</td>
<td>2</td>
<td>CLASSROOM</td>
<td>2</td>
<td>PRESENTLY</td>
</tr>
<tr>
<td>3</td>
<td>FEELING</td>
<td>2</td>
<td>FACTORS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>HOW</td>
<td>2</td>
<td>VIEWED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>BEYOND</td>
<td>2</td>
<td>POOR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>MUCH</td>
<td>2</td>
<td>BASED</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix B

Microcomputer-Implemented Content Analysis Activities
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 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. The activities are described in more detail in the next four sections.

Content Analysis Activity 1

The purposes of this activity were to process the information generated by the participants during Task 1 and prepare the materials for them to use during Task 2. Processing involved entering each participant's identification information, category identifiers, summaries, and response codes onto the microcomputer database.

Three documents were then prepared for each participant. Versions for fictitious participants are used in this appendix. The first document, Exhibit 1, was a personalized (addressed to the individual participant) form memo that was otherwise the same for all participants. It was used to provide the instructions for Task 2. The second document was different for each participant but it was prepared by the same process for all. It amounted to a typed version of the participant's Category Development Worksheet. An example is displayed as Exhibit 2. The third document, the Sample of Responses, was different for each participant and it was prepared one way for
MEMO

October 25, 1984

TO: Audry Farber
FROM: Dick Frisbie

SUBJECT: HPS Accountability Study

Thank you for working on developing a category system for coding responses to the open-ended question of the HPS Accountability Study. I have processed your categories and entered your codes for the sample of responses.

Please check over the materials I have returned to you, Audry. Make any changes you think are needed and return the final categories and codes for the sample of responses to me on November 1. Dr. Powerful has not yet returned from her meeting, but I would like to have these materials ready for her when she returns.

The documents for both groups had the codes given to the responses typed in a column next to each response with another column added to record a change of codes, if desired. In addition, the experimental group had their responses sorted into five groups arranged by the codes assigned to each response. The identifier and summary of the applicable category was also placed at the beginning of each group of responses. Exhibit 3 is an example. The control group had their responses in the same order as Task 1 with no category identifiers or summaries placed anywhere in the document.
Exhibit 2

Processed Category Development Worksheet

ISU Accountability Study for Hometown Public Schools

CATEGORY DEVELOPMENT WORKSHEET

ID: 0  NAME: Nick Danger

Cat # 1  Identifier: QUANTIFYING EFFECTIVE TEACHING
Summary: There are variables involved in classroom instruction which make it difficult to fairly evaluate all teachers by constant standards.

Cat # 2  Identifier: NEGATIVISM OF ADMINISTRATION
Summary: It is difficult to expect success for a controversial evaluation system if management cannot refrain from damaging teacher morale.

Cat # 3  Identifier: OBJECTIVITY OF PEER EVALUATION
Summary: Teachers will respond to a chance to evaluate one another by rating colleagues highly in return for high ratings.

Cat # 4  Identifier: COMPETITIVENESS AMONG STAFF
Summary: This system promotes unproductive rivalries and tensions among staff members who feel pressured.

Cat # 5  Identifier: SUPPORT CONCEPT, REJECT THIS DESIGN
Summary: Accountability systems are based on sound principles, but this one is not suitable.
Exhibit 3

Independent State University

USING A MAIL SURVEY TO ASSESS THE ACCOUNTABILITY SYSTEM
FROM THE PERSPECTIVE OF THE TEACHERS

SAMPLE OF RESPONSES TO THE OPEN-ENDED QUESTION

ID: 0 NAME: Professor Valery Powerful
TIME NEEDED TO VERIFY CATEGORIES AND RESPONSES: _______ HRS. _______ MIN.

QUESTION: Please give us any comments or recommendations you would like to make about any part of the Hometown Public Schools accountability system.

Category # Old New Response

Cat #1 Identifier: SOUND CONCEPT INAPPROPRIATELY IMPLEMENTED
Summary: Accountability is important and valuable but not as it has been devised for use in Hometown.

2 A _ Any of the components could have been utilized effectively had they been presented in a positive, professional manner.

4 A _ GOATs are nothing more than good organization which no one can argue against but the manner in which it was devised and implemented in Hometown leaves much to be desired.

7 A _ Accountability seems a good thing to me. Testing seems to be a good thing. But the way they are implemented and pushed on Hometown teachers is wrong.

14 A _ As I see the system as a whole, it is very good in design. However, it is not being used to upgrade the level of achievement, but rather to do just the opposite.

22 A _ I feel there should be some type of accountability system but none like we are presently using.

26 A _ Accountability can be a useful measurement tool. However, the system here will ultimately fail because of how it has been run.

31 A _ Accountability, when used in a positive manner, could be useful. When an accountability model like that in Hometown is used, this defeats the purpose of teaching in the classroom.

34 A _ I'm sure the system has some merit. However, there are many kinks which need to be ironed out.

37 A _ A good idea gone wrong because of dissention between the teaching staff and those in high administrative positions. As a result, the students and accountability system have become of little use to each other and unpleasantness has replaced harmony.

38 A _ The Hometown accountability system must be viewed in its totality and not just in the individual component parts of it. In toto it is oppressive and stifling.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Accountability is important, but not as a fear developing tool. I was among the four who resigned.

The accountability system is a good idea gone bad.

The accountability system falls short when measuring some of the most important facets in life - honesty, getting along with others, and learning to be a winner and loser gracefully, self control, etc.

Cat #2 Identifier: DIVISIVENESS AMONG INSTRUCTIONAL STAFF
Summary: Implementation of the system has created tension and division among instructional staff.

Competition is increasing for high scores on HAT tests and good ratings by principals.

Principal evaluation of teachers became a "report card comparison" among teachers causing jealousies, pickiness, and accusations of "browning." I have seen a once-unified staff become polarized and unhappy.

Teachers stay in room and do not share.

Unfortunately the end result of the accountability system has been the tension and division between teachers, rather than the progress and development of our students.

It seems to imply that we must be in competition with our colleagues in order to be good teachers.

A merit pay system would be a mistake as it would force even the good teachers to become concerned only about themselves. Each would be trying to outdo the other and thus would cause limited sharing and exchanges of ideas and materials among teachers.

Accountability here is backbiting, and dividing (as he wants) teachers.

Cat #3 Identifier: LACKS PROVISION FOR CONTEXT VARIABLE:
Summary: Student variability and other context variables are ignored in the system.

Someone who is in the classroom dealing with all types of kids, some who cannot read, some who hardly ever come to school, some who are in and out of jail, this teacher can see that, and the rigid accountability model that neglects the above mentioned problems is pure "B*******!"

There are too many variables that enter in to make it work.

There are too many variables in the educational system for accountability to work.

Anytime you deal with young adults many variables are involved. I do not feel we can force teachers to accept unstable variables to play a part in evaluating.
29 C The system—in no way—considers the various elements beyond testing that goes into the make-up of individual classes.

32 C No one wants to take low students in their room anymore because the principal will look at their scores and think they are poor teachers because their students scored lower—terrible!

35 C Our system presently seems to be under the illusion that we have total control of the educational processes for each child. WE are responsible! B***** — we are partially responsible but not over home and peer group.

39 C It doesn't take into consideration that some children have different socio-economic, emotional, and educational backgrounds and support from parents that keep them from learning.

41 C The accountability system has little or no provisions for low I.Q.s, drugs, liquor, sex, home problems, lack of interest in school by student and/or family, etc.—but teachers still have to produce!

43 C Under the accountability system all teachers are rated on same standards and all classes are expected to make 1 year's growth, even if records show that group has never shown 1 year's growth.

Cat #4 Identifier: COLLUSION ON PEER REVIEWS
Summary: Teachers deliberately give peers high ratings in order to protect each other.

1 D Peer ratings of a teacher in this system becomes an exercise of writing "5 for excellent."

11 D Peer ratings are a joke!! All teachers rate each other straight 5's—Excellent.

21 D Peer ratings: in my experience, have been done with the highest rating on each point.

24 D As to teacher peer ratings we have an agreement in our building that no one's rated lower than "good."

25 D We all got together in our school and rated each other No. 5 on the scale (excellent).

Cat #5 Identifier: SYSTEM INHUMANE
Summary: The system is viewed as lacking the human element and ignores human relations.

3 E The administration was quick to criticize, demand, and put pressure on us, but slow (if ever) to recognize, praise, and encourage us as human beings.

10 E The superintendent is too heavy handed and relies on threats when he wants to sell a program instead of working with us.

(page 4 omitted)
Content Analysis Activity 2

The purposes of this activity were to process the information generated by the participants during Task 2 and prepare the materials for them to use during Task 3. Relatively little work was required to process the information generated during Task 2. Only changes to existing identifiers, summaries, or codes for specific responses were made, if so indicated on a participant's Category Development Worksheet or Sample of Responses. Otherwise, the information existing on the database was retained. This concluded the procedures for the category experiment. All the remaining procedures apply only to the coding experiment.

Three documents were prepared for each participant to be used during Task 3. The first document was a form memo, individually addressed, that was otherwise the same for all participants. It was used to provide instructions for Task 3. The second document, the Official Categories Summary, Exhibit 4, was identical for all participants. It contained the final set of categories to be used by all participants to code responses for the remainder of the simulation. The third document, the Complete Set of Responses, was different for each participant, and it was prepared one way for experimental participants and another way for control participants. The Complete Set of Responses contained 100 responses, the 50 responses used during the first two tasks plus 50 new responses. The document was prepared the same way the Sample of Responses was prepared for each group during Content Analysis Activity 1. An individual participant's
Exhibit 4
Official Categories Summary

ISU Accountability Study for Hometown Public Schools
Dr. Valery Powerful, CENTER Director

OFFICIAL CATEGORIES SUMMARY

Cat. A Identifier: SOUND CONCEPT INAPPROPRIATELY IMPLEMENTED
Summary: Accountability is important and valuable but not as it has been devised for use in Hometown.

Cat. B Identifier: DIVISIVENESS AMONG INSTRUCTIONAL STAFF
Summary: Implementation of the system has created tension and division among instructional staff.

Cat. C Identifier: LACKS PROVISION FOR CONTEXT VARIABLES
Summary: Student variability and other context variables are ignored in the system.

Cat. D Identifier: COLLUSION ON PEER REVIEWS
Summary: Teachers deliberately give peers high ratings in order to protect each other.

Cat. E Identifier: SYSTEM INHUMANE
Summary: The system is viewed as lacking the human element and ignores human relations.

Complete Set of Responses looked different than the Sample of Responses to the extent that changes were made to specific response codes plus 50 uncoded responses were added after the 50 coded responses. The result was that the experimental participants had a Complete Set of Responses with 50 coded responses sorted by their own five categories and headed by their own identifiers and summaries. The 50 uncoded responses followed. The control participants had a Complete
Set of Responses with 50 coded responses in the same order as in Task 1 plus 50 uncoded responses added after the first 50.

**Content Analysis Activity 3**

The purposes of this activity were to process the information generated by the participants during Task 3 and prepare the materials for them to use during Task 4. Processing only involved the codes for the responses, and all 100 responses were processed for each participant. For the first 50 responses, numeric codes were replaced by alphabetic codes. For the second 50 responses, previously uncoded responses were given alphabetic codes.

Three documents were prepared for each participant to be used during Task 4. The first document was a form memo, individually addressed, that was otherwise the same for all participants. It was used to provide instructions for Task 4. The second document was another Official Categories Summary. It was provided to ensure each participant had a copy available for the task. The third document was an updated Complete Set of Responses. It was different for each participant, and it was prepared one way for experimental participants and another way for control participants. For experimental participants, the 100 coded responses were sorted by the five Dr. Powerful (Response Panel) categories, and each group of responses was headed by the appropriate Dr. Powerful identifier and summary. For the control participants, the 100 coded responses printed in the same order as they were for Task 3, and no category identifiers or summaries were printed anywhere on the document.
Content Analysis Activity 4

The purpose of this activity was to process the information generated by the participants during Task 4. Relatively little work was required to process this information. Only changes to existing codes for specific responses were made, if so indicated on a participant's Complete Set of Responses. Otherwise, the information existing on the data base was retained. This concluded the procedures for the coding experiment.
Appendix C

Development of the Category Hierarchy
Development of the Category Hierarchy

The category hierarchy was created for two basic purposes. First, it provided the means for deriving the measures of two dependent variables, category reliability and category validity. Second, it provided a qualitative framework for characterizing the categories developed by the participants. 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 place in the framework.

Recruit a Panel

A group of three doctoral students from Western Michigan University's Department of Educational Leadership were recruited to serve on the Hierarchy Panel. All panelists had direct experience planning and conducting evaluation studies. They also had extensive work experience ranging from elementary schools to colleges.
Have Panelists Independently Classify the Pilot Study-Generated Categories

The task of the panelists was to build upon the basic classification framework used in the Hometown Public Schools simulation activity described in Read Me First (see Appendix D) in such a way that all categories generated during the pilot study could be assigned to a single position in the framework. The five categories created by the Response Panel were used as the initial categories in the framework. If broader, narrower, related, or completely unrelated categories needed to be added to the original framework, it was up to each panelist to do so.

The panelists were instructed to write identifier and summary pairs for each new category and to locate each new category in the existing structure by drawing a new representation of the hierarchy. They were also instructed to assign each pilot study category to the synonymous hierarchy category.

Written instructions were given to the panelists. They could also ask the researcher questions as needed. In addition, they received the Pilot Study-Generated Categories, a set of 45 categories developed by those participants, and a set of Hierarchy Development Worksheets similar to the participants' Category Development Worksheets (see Appendix B). The panelists were asked to perform the task in one week. However, the last packet was returned about two weeks later.
Process the Panelists' Work on the Pilot Study Categories

The researcher processed the panelists' categories and assignments of participants' categories to the hierarchy in order to facilitate the group activity to follow. Two new documents were produced for distribution at the meeting: (1) the Hierarchy Panel Pilot Study Category Summary, a listing of all the original and new categories of the hierarchy; and (2) the Hierarchy Panel Classifications of Pilot Study-Generated Categories, a listing of all categories generated by the pilot study participants with the panelists' corresponding classification codes.

This processing consisted of: (a) entering the panelists' category identifiers and summaries into a data base on a DECmate II, (b) mapping the panelists' categories onto each other by matching roughly synonymous summaries and then changing each panelist's numeric category label to an alphabetic label, (c) translating each panelist's numeric code for a participant's category to the corresponding alphabetic code for the category, (d) entering each panelist's alphabetic code and all the corresponding participant categories into another data base on the DECmate II, (e) sorting the panelists' category identifiers and summaries by the alphabetic labels and printing them as the Hierarchy Panel Pilot Study Category Summary, and (f) sorting the participants' categories by the most used panelist code and printing the categories and codes as the Hierarchy Panel Classifications of Pilot Study-Generated Categories.
Cooperatively Determine the Pilot Study Category Hierarchy

The three panelists and the researcher met to determine the final definitions and framework of the pilot study category hierarchy and to assign each participant category to its proper location in order to verify that the hierarchy could be used to classify all the categories. This group of four people constituted the hierarchy panel as far as decisions at the meeting were concerned. At least three out of four people had to agree before a decision was final.

The first task was to select new categories, finalize their identifiers and summaries, and locate them in the framework. Each panelist was given a copy of the Hierarchy Panel Pilot Study Category Summary, the Hierarchy Panel Classifications of Pilot Study-Generated Categories, and photocopies of each hierarchy framework developed by the panelists. Panelists discussed each category and selected the unique additions to the framework. The researcher wrote each identifier and summary pair on sheets of posted newsprint as each decision was made. The researcher also redrew the framework on a chalkboard as each new category was added. When this task was completed, each category in the framework was assigned an identification letter, and the framework was copied to a sheet of newsprint.

The second task was to assign the 45 participant-generated categories to their appropriate locations in the newly created framework. The group considered and recoded each participant category that was not unanimously coded into the same alphabetic hierarchy category by the three panelists. After completion of this task, the panel meeting ended.
Have Panelists Independently Classify the Experiment-Generated Categories

The task of the panel here was to build upon the basic classification framework they had developed from the previous activity in such a way that all categories generated during the experiment could be assigned to a single position in the framework. The process they were instructed to use was directly analogous to that used with the pilot study-generated categories. The basic differences were that they had a more complete framework to start with and they had a new set of categories to use.

The panelists again received: (a) written instructions; (b) the Experiment-Generated Categories, a set of 370 categories generated by 75 participants; and (c) the Hierarchy Development Worksheets. They were asked to perform the task in one week. However, because of the size of the task and the intervening Winter break, the last packet was returned about eight weeks later.

Process the Panelists’ Work on the Experiment Categories

This activity was directly analogous to processing the panelists’ work on the pilot study categories. The two documents produced for distribution at the forthcoming meeting were: (1) the Hierarchy Panel Experiment Category Summary, a listing of all the previous and new categories of the hierarchy; and (2) the Hierarchy Panel Classifications of Experiment-Generated Categories, a listing of all categories generated by the experiment participants with the panelists’ corresponding classification codes.
Cooperatively Determine the Experiment Category Hierarchy

The panelists and the researcher met to determine the final definitions and framework of the category hierarchy for the experiment and to assign each participant category to its proper location in the hierarchy. These assignments became the raw data for the dependent variables, category reliability and category validity. The procedures used were analogous to those used to cooperatively develop the pilot study hierarchy and classify those participant-generated categories. The basic differences were that the panel had the more complete hierarchy to start with and they had a new set of categories to classify. Two three-hour sessions, one week apart, were needed to create the final hierarchy and code the participant categories in relation to it.

The text of the final hierarchy is presented in Table 1. A graphic representation of the final hierarchy is presented in Figure 1. Categories G and M were initially created by the panel, but no participant categories were given these codes in the end. Seven hybrid categories were also created because many participant categories contained important features of two Hierarchy Panel categories.
<table>
<thead>
<tr>
<th>Category</th>
<th>Identifier</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>SOUND CONCEPT INAPPROPRIATELY IMPLEMENTED</td>
<td>Accountability is important and valuable but not as it has been devised for use in Hometown.</td>
</tr>
<tr>
<td>B</td>
<td>DIVISIVENESS AMONG INSTRUCTIONAL STAFF</td>
<td>Implementation of the system has created tension and division among instructional staff.</td>
</tr>
<tr>
<td>C</td>
<td>LACKS PROVISION FOR CONTEXT VARIABLES</td>
<td>Student variability and other context variables are ignored in the system.</td>
</tr>
<tr>
<td>D</td>
<td>COLLUSION ON PEER REVIEWS</td>
<td>Teachers deliberately give peers high ratings in order to protect each other.</td>
</tr>
<tr>
<td>E</td>
<td>SYSTEM INHUMANE</td>
<td>The system is viewed as lacking the human element and ignores human relations.</td>
</tr>
<tr>
<td>F</td>
<td>UNSOUND CONCEPT</td>
<td>Accountability is an unworkable concept.</td>
</tr>
<tr>
<td>G</td>
<td>BIASED PUBLIC COMMUNICATIONS—not used alone</td>
<td>The Accountability System promotes a biased view of school to the public.</td>
</tr>
<tr>
<td>H</td>
<td>ACCOUNTABILITY SYSTEM DETRACTS FROM INSTRUCTIONAL ACTIVITIES</td>
<td>Attention to system detracts from instructional and related activities.</td>
</tr>
<tr>
<td>I</td>
<td>ILL-DEFINED, NEGATIVE</td>
<td>Negative in tone but lacking in specificity, and uninterpretable.</td>
</tr>
<tr>
<td>J</td>
<td>NEUTRAL COMMENT</td>
<td>Comment reports about some aspect of the system in a neutral tone and without suggestion or implication.</td>
</tr>
<tr>
<td>K</td>
<td>ADMINISTRATION ABUSES TEACHERS</td>
<td>Administration is abusive to teachers.</td>
</tr>
<tr>
<td>L</td>
<td>PERFORMANCE APPRAISAL INVALID</td>
<td>The system for appraising performance is invalid.</td>
</tr>
</tbody>
</table>
Table 1 (continued)

<table>
<thead>
<tr>
<th>Category</th>
<th>Identifier</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category N</td>
<td>Identifier: SYSTEM DETRIMENTAL TO STUDENTS</td>
<td>The Accountability System is detrimental to student learning and progress.</td>
</tr>
<tr>
<td>Category AH</td>
<td>Identifier: A &amp; H</td>
<td>Categories A &amp; H Combined.</td>
</tr>
<tr>
<td>Category AB</td>
<td>Identifier: A &amp; B</td>
<td>Categories A &amp; B Combined.</td>
</tr>
<tr>
<td>Category BD</td>
<td>Identifier: B &amp; D</td>
<td>Categories B &amp; D Combined.</td>
</tr>
<tr>
<td>Category BE</td>
<td>Identifier: B &amp; E</td>
<td>Categories B &amp; E Combined.</td>
</tr>
<tr>
<td>Category BL</td>
<td>Identifier: B &amp; L</td>
<td>Categories B &amp; L Combined.</td>
</tr>
<tr>
<td>Category CE</td>
<td>Identifier: C &amp; E</td>
<td>Categories C &amp; E Combined.</td>
</tr>
<tr>
<td>Category GK</td>
<td>Identifier: G &amp; K</td>
<td>Categories G &amp; K Combined.</td>
</tr>
</tbody>
</table>
Figure 1. Final Category Hierarchy
Appendix D

Simulation Materials for Session 1
Simulation Materials for Session 1

The three purposes of the Session 1 were to: (1) introduce the study, (2) provide a classroom lecture on content analysis, and (3) start the simulation activity used as the organizer for the content analysis task to follow. The researcher conducted the session according to a detailed script (Exhibit 1) organized by the three purposes identified above. The script ensured a degree of consistency for Session 1 between classrooms. It was not distributed.

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 (Exhibit 2); (b) Read Me First (Exhibit 3); (c) a Draft Introduction of an evaluation report (Exhibit 4); (d) a Practice Exercise (Exhibit 5); (e) a Practice Exercise Answer Sheet (Exhibit 6); and (f) the final group of handouts in a sealed envelope (discussed in Appendices A and B). A detailed description of the first session in relation to the rest of the procedures is discussed in the Experiments section of Chapter 3.
I. INTRODUCTION

MY NAME IS DICK FRISBIE. I'M HERE TO ASK YOU TO PARTICIPATE IN A STUDY ABOUT USING THE OUTPUT FROM MICROCOMPUTER PROGRAMS TO HELP CONDUCT CONTENT ANALyses OF RESPONSES TO OPEN ENDED SURVEY QUESTIONS. THIS STUDY IS PART OF MY DOCTORAL DISSERTATION AT WESTERN MICHIGAN UNIVERSITY.

YOU ARE ABOUT TO PARTICIPATE IN AN EXPERIMENT TO DETERMINE IF THE OUTPUT FROM COMMERCIALY AVAILABLE MICROCOMPUTER PROGRAMS CAN HELP PEOPLE DO A BETTER JOB AT CONDUCTING A CONTENT ANALYSIS OF RESPONSES TO OPEN-ENDED SURVEY QUESTIONS. YOU WILL NOT BE USING ANY MICROCOMPUTERS YOURSELF, BUT ONLY SOME OF THE OUTPUTS THEY PRODUCE.

WHAT I WILL DO NEXT IS PROVIDE AN OVERVIEW OF THE SIMULATION, PRESENT A BRIEF LECTURE ON CONTENT ANALYSIS, AND THEN START THE SIMULATION ACTIVITY WHICH IS THE CORE OF THE STUDY.

THE SIMULATION PROBLEM IS BASED ON A NON-TRIVIAL, REAL-WORLD EVALUATION STUDY THAT USED A MAIL SURVEY OF PUBLIC SCHOOL TEACHERS TO ASSESS THEIR REACTIONS TO A CONTROVERSIAL ACCOUNTABILITY SYSTEM. THE ACTUAL MAIL SURVEY IS NOT DIRECTLY RELATED TO MY STUDY OF CONTENT ANALYSIS EXCEPT THAT IT PROVIDES AN INTERESTING AND REALISTIC SETTING FOR THE SIMULATION.

FOUR OUT-OF-CLASS TASKS MUST BE PERFORMED BY THE PARTICIPANTS IN THIS STUDY. YOU MUST FIRST DEVELOP A CATEGORY SYSTEM INTO WHICH THE RESPONSES TO THE OPEN-ENDED QUESTION CAN BE CODED. THIS CATEGORY SYSTEM SHOULD BE BASED ON THE BACKGROUND MATERIAL PROVIDED AND A SAMPLE OF ACTUAL RESPONSES. I WILL ASK YOU TO TAKE NO MORE THAN ABOUT ONE HOUR TO COMPLETE THIS TASK. A GROUP OF EXPERTS IN THIS FIELD COMPLETED A SIMILAR TASK WITH FOUR TIMES AS MANY ITEMS AS YOU WILL HAVE IN AN AVERAGE OF ONE TO TWO HOURS.

THE SECOND TASK WILL BE TO VERIFY THE SYSTEM PREVIOUSLY DEVELOPED. THIS IS A SORT OF "LAST CHANCE TO CHECK YOUR WORK" TASK. IT SHOULD TAKE NO MORE THAN ABOUT A HALF HOUR TO COMPLETE.

THE THIRD TASK WILL BE TO CODE A SET OF RESPONSES BASED ON AN EXISTING SET OF CATEGORIES. THESE CATEGORIES MAY OR MAY NOT BE THE SAME AS THE ONES YOU DEVELOP. IT SHOULD TAKE NO MORE THAN ABOUT ONE HOUR TO COMPLETE.

THE FOURTH AND FINAL TASK WILL BE TO VERIFY THE CODES GIVEN TO RESPONSES IN THE THIRD TASK—ANOTHER "LAST CHANCE TO CHECK YOUR WORK" TASK. IT SHOULD TAKE NO MORE THAN ABOUT A HALF HOUR TO COMPLETE.
FOR THE NEXT EIGHT WEEKS YOU WILL BE GIVEN A TASK EVERY OTHER WEEK. I WILL USE THE ALTERNATE WEEK TO PROCESS YOUR WORK AND PREPARE THE MATERIALS FOR YOUR NEXT TASK, IF ONE FollowS. AFTER EVERYONE HAS TURNED IN THEIR FINAL TASK, I WILL ATTEMPT TO ANSWER ANY QUESTIONS YOU MAY HAVE ABOUT THE STUDY.

PARTICIPATION IN THE STUDY IS COMPLETELY VOLUNTARY, BUT THE TASK IS INTERESTING AND RELATIVELY PAINLESS. IT PROVIDES AN EXCELLENT OPPORTUNITY TO LEARN AND PRACTICE CONTENT ANALYSIS TECHNIQUES ON A VERY COMMON TYPE OF INFORMATION COLLECTED BY HUMAN SERVICE ORGANIZATIONS—RESPONSES TO OPEN-ENDED SURVEY QUESTIONS. HOPEFULLY, PARTICIPATION IN THE STUDY WILL HELP YOU TO BECOME BETTER EVALUATORS OR RESEARCHERS WHEN THE NEED ARISES AND BETTER PARTICIPANTS WHEN ASKED TO OFFER YOUR INPUT AGAIN. IN ADDITION [YOUR INSTRUCTOR] WILL BE GIVING YOU SOME INCENTIVES TO PARTICIPATE IN THIS STUDY. FIRST, YOU WILL BE GIVEN MUCH ENCOURAGEMENT AND MORAL SUPPORT TO COMPLETE THE TASKS. IN ADDITION, [LIST CONTRACTUAL INCENTIVES].

ARE THERE ANY QUESTIONS BEFORE I BEGIN THE LECTURE ON CONTENT ANALYSIS?

II. LECTURE

THIS LECTURE ON CONTENT ANALYSIS WILL FOCUS ON PROVIDING ANSWERS TO FOUR PRACTICAL QUESTIONS ABOUT CONTENT ANALYSIS. THE QUESTIONS ARE: 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?

EACH QUESTION COULD BE ANSWERED DIFFERENTLY THAN THE ONES I WILL PROVIDE AND MANY OTHER RELEVANT QUESTIONS COULD BE ADDRESSED AS WELL; BUT THIS LECTURE WILL GIVE YOU A GOOD BACKGROUND ON SOME GENERAL CHARACTERISTICS OF CONTENT ANALYSIS AND POINT OUT HOW THE SIMULATION RELATES TO A BROADER UNDERSTANDING OF CONTENT ANALYSIS.

HERE IS A HANDOUT THAT SUMMARIZES MY ANSWERS TO THE FOUR QUESTIONS. I WON’T DISCUSS ALL OF THE INFORMATION ON THE HANDOUT NOW, BUT I HAVE PROVIDED IT FOR YOUR FUTURE REFERENCE. RIGHT NOW I’LL JUST POINT OUT HOW THE QUESTIONS AND ANSWERS ARE RELATED TO THE SIMULATION PROBLEM YOU ARE ABOUT TO BEGIN. [BO]

1. WHAT IS IT?

[READ DEFINITION.] [POINT OUT KEY ELEMENTS OF THE DEFINITION IN THE HANDOUT.]

THIS DEFINITION OF CONTENT ANALYSIS FITS VERY WELL WITH THE
SIMULATION BECAUSE IT IS DIRECTED YET FLEXIBLE ENOUGH TO SUPPORT THE TYPES OF VALUE-BASED ASSESSMENTS USUALLY FOUND IN AN EVALUATION STUDY SUCH AS THE ONE ON WHICH THE SIMULATION IS BASED. THIS SIMULATION IS TYPICAL OF MANY EVALUATION STUDIES CONDUCTED IN EVALUATION.

2. WHAT ARE ITS USES?

I HAVE PROVIDED SOME GENERAL USES OF CONTENT ANALYSIS ALONG WITH SOME MORE SPECIFIC EXAMPLES. THE SPECIFIC EXAMPLES INDICATE THE WIDE RANGE OF USES TO WHICH CONTENT ANALYSIS HAS BEEN PUT; BUT I EXPECT ANY ONE OF US COULD ADD AT LEAST A FEW MORE EXAMPLES TO THE LIST. IT IS A VERY VERSATILE TECHNIQUE.

GENERALLY SPEAKING, THE SIMULATION WILL USE CONTENT ANALYSIS TO DESCRIBE CHARACTERISTICS OF COMMUNICATION. SPECIFICALLY, IT WILL HAVE A COMBINED EMPHASIS AS AN AID IN TECHNICAL RESEARCH OPERATIONS (TO CODE RESPONSES TO AN OPEN-ENDED QUESTION IN A MAIL SURVEY) IN ORDER TO REFLECT THE ATTITUDES, INTERESTS, AND VALUES OF A GROUP OF PEOPLE (TEACHERS WORKING IN A PARTICULAR PUBLIC SCHOOL DISTRICT). THIS IS ONE OF THE MOST COMMON USES OF CONTENT ANALYSIS IN EDUCATIONAL EVALUATION.

3. WHEN CONDUCTING SURVEYS, WHEN SHOULD IT BE USED WITH OPEN-ENDED QUESTIONS, INSTEAD OF USING QUANTITATIVE ANALYSIS WITH FORCED-CHOICE QUESTIONS?

THE THIRD QUESTION AND ANSWER HAVE BEEN PROVIDED BECAUSE A COMMON USE OF CONTENT ANALYSIS IN SOCIAL SCIENCE AND HUMAN SERVICE ORGANIZATIONS IS TO ANALYZE RESPONSES TO OPEN-ENDED SURVEY OR INTERVIEW QUESTIONS. AS I JUST STATED, THE SIMULATION USES CONTENT ANALYSIS OF RESPONSES TO AN OPEN-ENDED MAIL SURVEY QUESTION.

THE CLASSIC COMPETITOR WITH CONTENT ANALYSIS OF RESPONSES TO OPEN-ENDED QUESTIONS IS SOME TYPE OF QUANTITATIVE ANALYSIS OF RESPONSES TO FORCED-CHOICE QUESTIONS. SUCH QUESTIONS HAVE A LIMITED AND KNOWN SET OF RESPONSES FROM WHICH A RESPONDENT MUST CHOOSE. THE LIBERAL, IF ANY, LIMITATIONS ON RESPONSES TO OPEN-ENDED QUESTIONS PERMIT AN ALMOST LIMITLESS NUMBER OF DIFFERENT RESPONSES.

MY ANSWER TO THIS QUESTION INCLUDES THE ADVANTAGES AND DISADVANTAGES OF BOTH FORCED-CHOICE AND OPEN-ENDED QUESTIONS. I ALSO SUGGEST UNDER WHAT CONDITIONS WHICH TYPE OF QUESTION AND ANALYSIS SHOULD BE USED. [READ RECOMMENDATIONS WITH EMPHASIS ON OR’S & AND.]

4. HOW IS IT DONE?

FINALLY, I ASK HOW CONTENT ANALYSIS IS DONE AND PROVIDE A GENERAL OUTLINE FOR HOW TO DO IT. MOST PEOPLE WILL NOT BE DIRECTLY INVOLVED
IN ALL OF THE STEPS LISTED HERE, THEY WILL USUALLY BE INVOLVED IN ONLY A FEW OF THE ACTIVITIES. YOU WILL SPEND MOST OF YOUR TIME DEVELOPING CODING INSTRUCTIONS BY CREATING A CATEGORY SYSTEM, AND YOU WILL TRANSFORM, OR CODE, DATA—THAT IS, ASSIGN CATEGORY CODE NUMBERS OR LETTERS TO INDIVIDUAL RESPONSES.

THE FINAL PAGE SHOWS SOME COMMON WAYS OF REPORTING THE RESULTS OF A CONTENT ANALYSIS. THIS EXAMPLE IS BASED ON ONE QUESTION A SCHOOL DISTRICT MIGHT ASK A SAMPLE OF REGISTERED VOTERS AS PART OF A COMMUNITY SURVEY. THE DISTRICT MAY PREFER TO HAVE THIS BE AN OPEN-ENDED QUESTION RATHER THAN A FORCED-CHOICE QUESTION IF THEY DON'T WANT TO "LEAD" THE VOTERS TO GIVE ANY PARTICULAR RESPONSES, EVEN IF THEY HAVE A GOOD IDEA IN ADVANCE HOW THE RESPONSES WILL BE PUT INTO CATEGORIES. THE PURPOSES OF THIS ANALYSIS WERE TO DESCRIBE THE VOTERS' RESPONSES TO THE QUESTION THIS YEAR AND COMPARE THEM TO HOW THE VOTERS RESPONDED TO THE SAME QUESTION LAST YEAR.

THE FIRST SECTION REPRESENTS A NARRATIVE APPROACH TO SUMMARIZING THE RESPONSES. TYPICAL RESPONSES ARE PRESENTED FOR EACH CATEGORY WITH SOME SUMMARY STATISTICS. A LONGER VERSION OF THIS SECTION WOULD INCLUDE ELABORATED DEFINITIONS OF THE CATEGORIES, A FEW MORE ACTUAL RESPONSES, AND A SYNTHESIS OF AN OVERALL IMPRESSION OF THE GROUP OF RESPONSES.


THE THIRD SECTION REPRESENTS A GRAPHIC APPROACH TO DESCRIBING THE RESPONSES. THE FREQUENCIES FROM THE FIRST TABLE WERE CONVERTED TO PROPORTIONS AND PERCENTS THEN PRESENTED IN SIDE-BY-SIDE BAR GRAPHS. THIS PRESENTATION MAKES IT RATHER EASY TO NOTICE THE RELATIVELY LARGE DROP IN CONCERN FOR FINANCES AND RISE IN CONCERN FOR THE 3 R'S INDICATED IN THE CHI SQUARE ANALYSIS. IT ALSO MAKES IT EASIER TO NOTICE THE MOST IMPORTANT ISSUE IN 1983 WAS FINANCES, BUT THE MOST IMPORTANT ISSUE IN 1984 WAS DISCIPLINE.

ARE THERE ANY QUESTIONS BEFORE WE GO ON TO THE SIMULATION?
III. SIMULATION

THIS SIMULATION IS BASED ON AN ACTUAL EVALUATION STUDY CONDUCTED BY A GROUP OF INDEPENDENT EVALUATORS FOR A TEACHERS' UNION OF A PUBLIC SCHOOL DISTRICT. THE UNION AND THE ADMINISTRATION WERE HOTLY DIVIDED OVER THE DISTRICT'S ACCOUNTABILITY SYSTEM. THE PURPOSE OF THE EVALUATION STUDY WAS TO ASSESS THE ACCOUNTABILITY SYSTEM FROM THE PERSPECTIVE OF THE TEACHERS. THE PURPOSE OF THIS SIMULATION IS TO PROVIDE AN INTERESTING BUT REALISTIC BACKDROP TO TEST SOME CONTENT ANALYSIS TECHNIQUES IMPLEMENTED WITH THE USE OF OUTPUT FROM MICROCOMPUTER PROGRAMS. THE SIMULATION COULD HAVE BEEN BASED ON OTHER STUDIES IN OTHER FIELDS AND STILL BE USED TO TEST THE CONTENT ANALYSIS TECHNIQUES OF INTEREST.

READ ME FIRST

THE SIMULATION BEGINS WITH READ ME FIRST. PLEASE LOOK OVER THIS HANDOUT AS I EXPLAIN IT TO YOU. [HAND OUT] READ ME FIRST SETS THE STAGE FOR THE SIMULATION AND GIVES YOU BASIC INSTRUCTIONS FOR THE FIRST TASK. THE FIRST PAGE SETS UP THE SIMULATION. YOU ARE NOW A STUDENT RESEARCHER WHO MUST FILL IN FOR THE DIRECTOR OF THE CENTER AT ISU, DR. POWERFUL. YOUR JOB IS TO CREATE A FIVE CATEGORY CLASSIFICATION SYSTEM CONSISTENT WITH THE GENERAL FRAMEWORK PICTURED IN FIGURE 1 ON THE THIRD PAGE. YOUR INSTRUCTIONS ARE ON PAGE 2. FIRST, YOU WILL READ SOME MORE INFORMATION HERE AND COMPLETE A PRACTICE EXERCISE. THEN YOU WILL BE GIVEN A PACKET OF MATERIALS LIKE THE PRACTICE EXERCISE TO WORK ON AT HOME. BASICALLY, YOUR JOB IS TO CREATE FIVE NEW CATEGORIES, CODE THE RESPONSES ACCORDING TO THOSE CATEGORIES, AND RETURN YOUR WORK TO ME NEXT WEEK. PLEASE READ ALL OF THE DETAILS IN THE INSTRUCTIONS RIGHT BEFORE YOU BEGIN YOUR WORK AT HOME.

DRAFT INTRODUCTION

HERE IS THE DRAFT INTRODUCTION OF THE EVALUATION REPORT PREPARED BY DR. POWERFUL. IT CONTAINS A "BACKGROUND OF THE PROBLEM" AS SEEN BY DR. POWERFUL. IT IS INTENDED TO GIVE YOU A FEEL FOR THE CONTEXT IN WHICH THE MAIL SURVEY WAS CONDUCTED. PLEASE TAKE A MINUTE TO LOOK AT IT NOW AND SPEAK OUT IF YOU HAVE ANY QUESTIONS. [HAND OUT]

PRACTICE EXERCISE

HERE IS A PRACTICE EXERCISE DEVELOPED BY DR. POWERFUL. IT SHOULD GIVE YOU AN IDEA ABOUT WHAT KINDS OF CATEGORY IDENTIFIERS AND SUMMARIES SHE WANTS YOU TO WRITE. AFTER EVERYONE HAS COMPLETED THE EXERCISE, I WILL HAND OUT AN ANSWER KEY AND WE WILL DISCUSS THE EXERCISE. FEEL FREE TO ASK QUESTIONS AS YOU GO ALONG. [HAND OUT]
PRACTICE EXERCISE ANSWER KEY

HERE IS THE ANSWER KEY TO THE EXERCISE. THE ANSWERS ARE PRINTED IN BOLDFACE TYPE. READ THE SUGGESTED ANSWERS AND LET ME KNOW IF YOU HAVE ANY QUESTIONS. [HAND OUT]

TASK 1 PACKET

THE PACKETS I AM ABOUT TO HAND OUT INCLUDE A TEAR-OFF IDENTIFICATION SHEET AND THE SET OF MATERIALS FOR YOUR FIRST TASK. PLEASE!! DO NOT OPEN YOUR PACKETS AROUND ANY OF YOUR CLASSMATES AND PLEASE DO NOT DISCUSS THE DETAILS OF YOUR ACTIVITIES RELATED TO THE SIMULATION WITH ANY OF YOUR CLASSMATES UNTIL EVERYONE HAS COMPLETED THE FOURTH TASK.

THE INFORMATION ON THE ID SHEETS WILL BE HELD STRICTLY CONFIDENTIAL. IT WILL ONLY BE USED FOR DATA IDENTIFICATION, COMPARING DEMOGRAPHIC INFORMATION BETWEEN GROUPS OF PARTICIPANTS, OR FOR ME TO CONTACT YOU IF THAT BECOMES NECESSARY LATER ON. PLEASE FILL OUT THE ID SHEET NOW AND PASS IT UP TO ME. PRINT YOUR NAME HOW YOU PREFER TO BE CALLED, GIVE A PHONE NUMBER WHERE YOU CAN MOST EASILY BE REACHED, AND FILL IN THE REMAINDER OF THE FORM TO THE BEST OF YOUR KNOWLEDGE OR JUDGMENT. [HAND OUT]

FROM NOW ON, EACH OF YOU WILL HAVE A SLIGHTLY DIFFERENT SIMULATION EXPERIENCE, BASED ON THE MATERIALS YOU RECEIVE AND ON HOW YOU PERFORM EACH OF THE FOUR TASKS. BECAUSE THIS STUDY IS INTENDED TO TEST A SET OF SPECIFIC RESEARCH HYPOTHESES, SHARING YOUR MATERIALS OR EXPERIENCES WITH OTHER PARTICIPANTS FROM NOW UNTIL THE FOURTH TASK IS COMPLETED WOULD THREATEN THE VALIDITY OF THE FINDINGS. AFTER EVERYONE HAS TURNED IN THEIR LAST SET OF MATERIALS, YOU SHOULD FEEL FREE TO DISCUSS THE STUDY WITH EACH OTHER. I WILL ALSO TRY TO ANSWER ANY QUESTIONS I CAN AFTER THAT TIME.

THAT’S ALL WE HAVE TO DO FOR NOW. REMEMBER, YOUR FIRST TASK IS TO DEVELOP A FIVE CATEGORY SYSTEM TO CODE THE RESPONSES TO THE OPEN-ENDED QUESTION. YOUR BASIC INSTRUCTIONS ARE ON THE SECOND PAGE OF READ ME FIRST. ITEM 12 LISTS THE TWO THINGS YOU NEED TO BRING BACK TO ME AT THE NEXT CLASS MEETING ON:

1) THE CATEGORY DEVELOPMENT WORKSHEET WITH ONE IDENTIFIER AND ONE SUMMARY FOR EACH OF THE FIVE CATEGORIES YOU DEVELOP; AND 2) THE SAMPLE OF RESPONSES WITH EACH RESPONSE CODED INTO ONLY ONE OF YOUR FIVE CATEGORIES.

THANKS FOR YOUR HELP WITH THIS STUDY. I’LL SEE YOU AT THE NEXT CLASS MEETING!
Exhibit 2
Content Analysis
Answers to Four Practical Questions

1. What is it?

"Content analysis is a research technique for making replicable and valid inferences from data to their context" (Krippendorff, 1980).

The main purpose of using content analysis is to make inferences—draw conclusions—from a set of data, such as, documents, conversations, survey responses, paintings, or photographs, to their context—the original source and situation in which the data were created. As a research technique, content analysis has its own special procedures for processing information. These procedures are designed to encourage inferences made to be both replicable—reliable, or reproducible by different researchers at different times and places—and valid—justifiable in terms of a well-defined, symbolic interpretation of the data.

2. What are its uses?

General Uses (Holsti, 1969)

* to describe characteristics of communication—asking what, how, and to whom something is said
* to make inferences as to the antecedents of communication—asking why something is said
* to make inferences as to the effects of communication—asking with what effects something is said

Specific Examples (Berelson, 1952)

* to describe trends in communication content
* to trace the development of scholarship
* to disclose international differences in communication content
* to compare media or "levels" of communication
* to audit communication against objectives
* to construct and apply communication standards
* to aid in technical research operations (to code open-ended questions in survey interviews)
* to expose propaganda techniques
* to measure the "readability" of communication materials
* to discover stylistic features
* to identify the intentions and other characteristics of the communicators
* to determine the psychological state of persons or groups
* to detect the existence of propaganda (primarily for legal purposes)
* to secure political and military intelligence
* to reflect attitudes, interests and values ("cultural patterns") of population groups
* to reveal the focus of attention
* to describe attitudinal and behavioral responses to communications

3. When conducting surveys, when should it be used with open-ended questions, instead of using quantitative analysis with forced-choice questions?

**Advantages, Disadvantages, & Guidelines for Choosing Between Forced-Choice and Open-Ended Questions** (Demaline & Quinn, 1979)

**Forced-Choice Questions**

**Advantages**
1. It is easier for respondent to answer.
2. Focuses respondent's answer on issues and data of importance to you. Respondents categorize themselves instead of you categorizing them.
3. More questions can be asked because time is saved by the respondent simply checking.
4. Precoded answers are easily analyzed.

**Disadvantages**
1. It requires advance information about possible response categories that may be given.
2. It may bias responses by suggesting answers.
3. It does not allow for diversity and richness in individual expression.

**Open-Ended Questions**

**Advantages**
1. It can easily be formulated without knowing the full range of answers that may be given.
2. It can accommodate questions for which a wide range of different answers will be given.
3. It does not condition or bias the answer as much as the forced-choice question.

**Disadvantages**
1. It requires the respondent to write a lot. Communication skills may influence the answer, in addition [to] the other respondent characteristics.
2. The respondent may address different facets of the question in which you may not be interested or may not give complete information in answering the question.
3. Fewer questions can be asked in a questionnaire because answers may be lengthy and time consuming to give.
4. Responses are difficult to analyze. The investigator must devise a coding scheme and then categorize responses based on this scheme. The diversity and richness of responses are usually reduced by this process, and it is time consuming.

**Use Forced-Choice Questions & Quantitative Analysis If**
1. it is important that the effort or verbal skills of the respondents be kept relatively low, OR
2. there is a clear understanding of what the likely or important responses will be, OR
3. a large number of questions need to be asked in relation to the time available, OR
4. it is important that the responses be easy to code for analysis.

**Use Open-Ended Questions & Content Analysis If**
1. the full range of likely and valid responses is not known or a wide range of responses is expected, OR
2. there is concern for biasing respondents if a set of possible answers is given, AND
3. the skills and time needed for coding potentially complex responses are available.

4. **How is it done?**

**Steps in Content Analysis** (Krippendorff, 1980)

**Design** (aspects)
* applying the framework for content analysis
* searching for suitable data
* searching for contextual knowledge
* developing plans for unitizing and sampling
* developing coding instructions
* searching for contextually justifiable procedures
* deciding on qualitative standards
* budgeting and resource allocation

**Execution** (contains one or more of the following)
* sampling by sampling units until the sample can be judged sufficiently representative of the population
* identification and description of recording units which must be reproducible and satisfy criteria of semantical validity where applicable
* data reduction and transformation of data into a form required for analysis, retaining all relevant information
* application of context-sensitive analytical procedures (analytical constructs) to yield inferences
* analysis, identification of pattern within inferences, testing hypotheses regarding relations between inferences, and results obtained by the methods and pragmatic validation of findings

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Exhibit 2 (continued)

Report (should be specific about some or all of the following)
* a statement of the general problem to which the research pertains
* an account of the background of the problem
* a statement of the specific objectives of the content analysis
* a justification of the choice of data, methods and design
* a description of the procedures actually followed
* a presentation of the findings
* a self-critical appraisal of the procedures followed and the results obtained

References


Exhibit 2 (continued)

A Comparison of 1983 vs. 1984 Responses to the Question: What is the most important issue facing the schools today?

<table>
<thead>
<tr>
<th></th>
<th>1983</th>
<th>1984</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Finances</td>
<td>39%</td>
<td>32%</td>
</tr>
<tr>
<td>e.g., The cost of a good education is getting too high.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. Discipline</td>
<td>36%</td>
<td>39%</td>
</tr>
<tr>
<td>e.g., Kids don't take responsibility for their actions any more.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. 3 R's</td>
<td>10%</td>
<td>16%</td>
</tr>
<tr>
<td>e.g., It's time to get back to the basics.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D. Other</td>
<td>15%</td>
<td>13%</td>
</tr>
<tr>
<td>e.g., You have no right to close the school in our neighborhood! We need fewer administrators and more teachers.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Most Important School Issue?
(Number of Voters)

<table>
<thead>
<tr>
<th></th>
<th>Finances</th>
<th>Discipline</th>
<th>3 R's</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1983</td>
<td>279</td>
<td>256</td>
<td>70</td>
<td>104</td>
<td>709</td>
</tr>
<tr>
<td>1984</td>
<td>231</td>
<td>281</td>
<td>89</td>
<td>111</td>
<td>712</td>
</tr>
</tbody>
</table>

(Contributions to Value of Chi Square)

<table>
<thead>
<tr>
<th></th>
<th>1983</th>
<th>1984</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finances</td>
<td>2.36</td>
<td>2.61</td>
<td>4.97*</td>
</tr>
<tr>
<td>Discipline</td>
<td>0.56</td>
<td>0.51</td>
<td>1.07</td>
</tr>
<tr>
<td>3 R's</td>
<td>5.89</td>
<td>3.72</td>
<td>9.61*</td>
</tr>
<tr>
<td>Other</td>
<td>0.57</td>
<td>0.67</td>
<td>1.24</td>
</tr>
<tr>
<td>Total</td>
<td>4.97*</td>
<td>1.07</td>
<td>9.61*</td>
</tr>
</tbody>
</table>

* Largest contributions to value of Chi Square sig. at alpha = 0.01

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
You are a student employee at the Center for Evaluating New Techniques in Educational Research (CENTER) at Independent State University (ISU). Your job is to assist the director, Dr. Valery Powerful, in whatever tasks she needs to accomplish. Because you have shown exceptional talent in past performances, Dr. Powerful has much confidence in your abilities—often giving you a chance to work on very difficult assignments.

The newest project at the CENTER involves assessing a controversial accountability system of a moderately large public school district in a nearby state. This project is jointly funded by the local, state, and national education associations. Because of several setbacks in devising a design agreeable to all parties concerned, particularly the superintendent, Dr. Powerful and the sponsors finally decided to use teacher responses to a mail questionnaire as the major data source. Consequently, this study has become an attempt to review the accountability system from the perspective of the teachers.

So far, you have not worked on this project because of your extensive involvement in another project, A Content Analysis of Homework Assignments for Introductory Research Courses at the Graduate Level. However, Dr. Powerful has just been called away for four weeks to attend an emergency meeting of the National Network of CENTER Directors in Honolulu. Before she went to the meeting, she put together some materials she was working on and left you this note:

Please take over the work on this accountability study in my absence. All of the questionnaires have been returned. Responses to the forced-choice questions have already been coded so don’t worry about those. I would like you to spend your time working on the responses to the open-ended questions. My secretary has already typed a draft of my introduction to the report, a sample of the responses to the open-ended question, and some worksheets I had planned to use.

I should warn you that the responses I have read so far have been overwhelmingly negative. This means that categories like "Negative Comments About the Accountability System" or "Negative Comments About the Superintendent" would not be particularly helpful. Instead, I want you to create a set of five categories, each of which captures a particular set of negative comments about the accountability system. These categories should be a) exhaustive—all
categories fit somewhere, b) mutually exclusive—there is one best category for each response, and c) consistent with the general framework in Figure 1—all categories are at the same level of detail and they all represent one group of similar negative comments about the accountability system. I made up a practice exercise for you to do before you start working on the actual responses. This should give you an idea of the level of detail and tone I want the categories to capture. DO NOT use any of these categories when you make up five of your own. My secretary will help you with the practice exercise when you are ready to begin.

Here is how I would like you to proceed:

1. READ the DRAFT INTRODUCTION of the report to get a flavor of the study.

2. COMPLETE the PRACTICE EXERCISE.

3. GO HOME and work on this ALONE! I want an INDEPENDENTLY DEVELOPED CATEGORY SYSTEM. If you have any questions about how to work on this assignment, call my secretary, Dick Frisbie, at 383-8166 (work) or 375-4271 (home). I had a chance to talk to him about how I wanted this done before I left for the emergency meeting.

4. FIND A LARGE, CLEAR WORK AREA TO SPREAD OUT YOUR MATERIALS.

5. READ the SAMPLE OF RESPONSES to the question.

6. CREATE EXACTLY FIVE CATEGORIES, no more, no less, INTO WHICH ALL RESPONSES MAY BE CODED. [Powerful people are sometimes arbitrary.]

7. USE A "GARBAGE" CATEGORY like "Other" or "Miscellaneous" IF YOU DECIDE ONE IS NECESSARY.

8. Use whatever techniques you think would work best for deciding what the five categories should be, except, DO NOT CUT UP THE CATEGORY DEVELOPMENT WORKSHEETS or SAMPLE OF RESPONSES. My secretary needs to use them after you complete your task.

9. EACH CATEGORY MUST HAVE a) a NUMBER—use 1 through 5, b) an IDENTIFIER—a brief, descriptive title for the category which indicates its negative nature, and c) a SUMMARY—an
operational definition of the category complete enough to allow people besides yourself to decide whether a particular response should or should not be included in that category.

10. **Write one identifier and summary for each number on the category development worksheets.**

11. **Put** the appropriate category number next to each response on the sample of responses.

12. **Return** the category development worksheets and the summary of responses to my secretary on July 19, 1984. He will then process your work and return it to you for a final review before we start coding all of the responses.

13. Because my secretary will return your work to you after he has processed it, **do not spend more than about one hour to complete this task.** You will get another chance to review your work and make any changes you think are necessary.

![Figure 1. A general framework for developing a category system to code responses to the open-ended question, "Please give us any comments or recommendations you would like to make about any aspect of the Hometown Public Schools accountability system."

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Exhibit 4
Draft Introduction

The CENTER, Independent State University

USING A MAIL SURVEY TO ASSESS THE ACCOUNTABILITY SYSTEM FROM THE PERSPECTIVE OF THE TEACHERS

Introduction

The report which follows is an evaluation of the Hometown Public Schools accountability system from the point of view of the teachers. At the time of this study the Hometown accountability system had been operating for four years under the leadership of the current superintendent. In a report issued by the School Administration two years ago the following background was provided:

Hometown Public Schools, in an effort to plan logically for improving academic achievement of all students and, at the same time, increase efficiency of school operation, organized a Department of Research and Development one school year ago.

Hometown students in grades one through nine were tested in September and May one school year ago, using the nationally standardized Hometown Achievement Tests (HAT). Students in all grades are being tested during the current school year.

Teachers receive test scores of individual students, composite scores for their class groups, and an analysis showing which items on the test were most frequently missed by their students.

Teachers also administer their own Goals & Objectives for Achievement on Teacher-made (GOAT) tests to determine how well students are learning the classroom material and/or skills.

This continual flow of information helps teachers to monitor their own performance and guides them as they plan lessons to move each student toward greater achievement and the conquering of weaknesses which inhibit that achievement.
The salaries of all Hometown school administrators, including the superintendent, are based on annual performance evaluations.

At the beginning of each school year, administrators develop performance objectives. At the end of the school year, a "percentage of accomplishment" is determined and ratings by several relevant groups are considered.

Performance profiles are developed for all teachers. The profile, which is shared with the teacher as a source of information for improving teacher performance, includes views of students, the principal, other teachers, parents, and the teacher's self-evaluation plus extensive student achievement data.

Information and comments are gathered, interpreted, and reported with the help of Research and Development staff members. The collected information is used to guide school personnel toward improving their performance.

The Hometown accountability system has received considerable national attention. For example, the Journal of American School Boards reported last year that by this summer "Hometown schools will probably have one of the most comprehensive computerized systems of personnel evaluation and accountability yet devised" (p. 34).

On the other hand, conflict over the school system's accountability program has been high for several years; charges and countercharges have been exchanged regularly between the Hometown Education Association (HEA) and the Office of Superintendent of Schools. The HEA, for example, has charged that teachers were being demoralized; the superintendent has argued that teachers don't want to be accountable. Politically charged statements have flowed with increasing frequency from both sources since the beginning of the accountability system—making constructive dialog increasingly more difficult.

This spring the HEA sought assistance from the state and national education associations to engage in a review of the accountability system. Because repeated attempts to enlist the participation of the Hometown Public Schools Administration have failed, this study has used as its major data source teacher responses to a mail questionnaire administered during the first week of June.

As the conflict between the administration and teachers has grown (including unsigned contracts, lawsuits, and countersuits), the HEA has asserted that the accountability system as implemented in
Hometown is destructive to the educational system. This survey of teacher attitudes about the experiences with the accountability system was undertaken in large measure to test that assertion. Therefore, the report which follows is an evaluation of the Hometown Public Schools accountability system from the perspective of the teachers.

Note. Blind adaptation from Patton et al. (1976).
ISU Accountability Study for Hometown Public Schools

PRACTICE EXERCISE

1. Using the sample responses as input, write an Identifier and two Summaries to complete the category system.
2. Using the category system you completed, assign category code numbers to the responses which do not yet have a code.

Categories

Category #1. Identifier: Teaching to the Test
Summary: Teachers drill their students with the actual test questions and answers prior to administration of the test.

Category #2. Identifier: Reprisals Against Teachers
Summary: ________________________________________________________________

Category #3. Identifier: _______________________
Summary: ________________________________________________________________

Responses

<table>
<thead>
<tr>
<th>No.</th>
<th>Cat.</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>___</td>
<td>If a teacher voices an opinion that is in disagreement with the administration, the teacher is subject to transfer and is given the most difficult students to try to teach. This has happened to me.</td>
</tr>
<tr>
<td>2.</td>
<td>___</td>
<td>We have lost focus on what kids in Hometown need to learn, and are teaching to a test.</td>
</tr>
<tr>
<td>3.</td>
<td>___</td>
<td>If you disagree, don’t say so or you will be reprimanded.</td>
</tr>
</tbody>
</table>
4. ___ It is my opinion that Hometown uses the "Accountability" program that they do in order to present a front to the public.

5. ___ As an untenured teacher I feel that my job is in jeopardy if I voice my opinion contrary to policy or the superintendent's feelings.

6. ___ Some teachers start to rehearse their students in January for the spring testing in May.

7. ___ This entire accountability system is a "game" designed to impress the community.

8. ___ The worst thing about accountability systems is that you find yourself teaching to the test.

9. ___ As you may well be aware, the program has resulted in a situation in which a huge amount of time and effort is spent in a "whitewash" attempt to appear highly successful.
Exhibit 6
Practice Exercise Answer Sheet

ISU Accountability Study for Hometown Public Schools

PRACTICE EXERCISE: ANSWER SHEET

1. Using the sample responses as input, write an Identifier and two Summaries to complete the category system.
2. Using the category system you completed, assign category code numbers to the responses which do not yet have a code.

Categories

Category #1. Identifier: Teaching to the Test
Summary: Teachers drill their students with the actual test questions and answers prior to administration of the test.

Category #2. Identifier: Reprisals Against Teachers
Summary: Teacher challenges to individual administrators or the accountability system in general are met with involuntary transfers, reprimands, dismissals, and other types of reprisals—real or anticipated.

Category #3. Identifier: Administration Misleading the Public
Summary: The administration deliberately misrepresents the implementation process and effects of the accountability system.

Responses

No. Cat. Response

1. 2 If a teacher voices an opinion that is in disagreement with the administration, the teacher is subject to transfer and is given the most difficult students to try to teach. This has happened to me.

2. 1 We have lost focus on what kids in Hometown need to learn, and are teaching to a test.

3. 2 If you disagree, don’t say so or you will be reprimanded.
4. It is my opinion that Hometown uses the "Accountability" program that they do in order to present a front to the public.

5. As an untenured teacher I feel that my job is in jeopardy if I voice my opinion contrary to policy or the superintendent's feelings.

6. Some teachers start to rehearse their students in January for the spring testing in May.

7. This entire accountability system is a "game" designed to impress the community.

8. The worst thing about accountability systems is that you find yourself teaching to the test.

9. As you may well be aware, the program has resulted in a situation in which a huge amount of time and effort is spent in a "whitewash" attempt to appear highly successful.
Appendix E

Researcher/Instructor Contract
AGREEMENT TO PARTICIPATE IN A CONTENT ANALYSIS STUDY

PARTIES TO THE AGREEMENT

This agreement is between Richard D. Frisbie, hereinafter referred to as the Researcher, and , hereinafter referred to as the Instructor. The Instructor agrees to participate in a content analysis study conducted during the Fall Semester, 1984, at Western Michigan University, by the Researcher. Instructor participation includes permitting the Researcher to recruit the students, hereinafter referred to as the Participants, enrolled in the Instructor’s class, , conducted during Fall Semester, 1984, at WMU.

OVERVIEW OF THE STUDY

The purpose of the study is to advance the body of knowledge concerning the extent to which computer outputs, produced from commercially available general purpose programs for microcomputers, and based on techniques for survey and discovery in content analysis, can be used to improve the reliability and validity of content analyses conducted on responses to open-ended questions found in evaluation studies. A set of four research hypotheses has been developed to assert that such techniques will indeed improve both the reliability and validity of category system development and coding responses in relation to an existing category system. The experimental study the Instructor hereto agrees to participate in is designed to test these hypotheses.

The experiment will use a simulation problem which is based on a non-trivial, real-world evaluation study. The study used a mail questionnaire to obtain the reactions of teachers to a controversial accountability system of a public school system in the Midwest. Some of the information obtained from the teachers was in the form of responses to a set of open-ended questions. The focus of the experiment will be to determine if certain types of computer output related to these responses will help experimental participants produce more reliable and valid category systems as well as, code responses with more reliability and validity, when compared to control participants.

SUMMARY OF THE PROCEDURES

The Participants of the experiment will receive a classroom lecture about the theories and methods of content analysis from the Researcher. The Participants then must independently perform a content analysis of 100 responses in four parts: (1) develop a
category system; (2) verify the system; (3) code a set of responses; and (4) verify the codes. The experiment will require eight classroom sessions to allow sufficient time for the Researcher and Participants to perform their respective tasks.

The schedule of group sessions (Sn) and Participants' out-of-class tasks (Tn) are summarized below:

<table>
<thead>
<tr>
<th>Date</th>
<th>Summary</th>
<th>(Time Estimate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Lecture, Setup, Ss receive 1st task materials (1 hr)</td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>Prticipants perform 1st task: develop a category system (1 hr)</td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td>Participants return 1st task materials (10 min)</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>Participants receive 2nd task materials (10 min)</td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td>Prticipants perform 2nd task: verify category system (1/2 hr)</td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>Participants return 2nd task materials (10 min)</td>
<td></td>
</tr>
<tr>
<td>S5</td>
<td>Participants receive 3rd task materials (10 min)</td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td>Participants perform 3rd task: code a set of responses (1 hr)</td>
<td></td>
</tr>
<tr>
<td>S6</td>
<td>Participants return 3rd task materials (10 min)</td>
<td></td>
</tr>
<tr>
<td>S7</td>
<td>Participants receive 4th task materials (10 min)</td>
<td></td>
</tr>
<tr>
<td>T4</td>
<td>Participants perform 4th task: verify the codes (1/2 hr)</td>
<td></td>
</tr>
<tr>
<td>S8</td>
<td>Participants return 4th task materials (10 min)</td>
<td></td>
</tr>
</tbody>
</table>

The Researcher also agrees to provide an in-class debriefing on the study on ______. Because of the timing of the debriefing, only preliminary findings will be presented.

**NON-DISCLOSURE**

Because the study uses a "double-blind" design, the Instructor agrees to refrain from (a) seeking information about assignments to treatments or (b) disclosing to the Participants the details of the problem, research design, variables, assignments to treatments, procedures, analyses, or any other information which would jeopardize the integrity or validity of the experiment until after all data is collected. The Instructor will also discourage Participants from discussing any details of the experiment until all data has been collected.

In addition, the Instructor agrees to refrain from distributing any descriptions of the study or its findings: (a) at public meetings, such as, conferences, workshops, or conventions; or (b) in written forms, such as, personal communications, monographs, or sections in newsletters, magazines, journals, or books; until after the successful defense of the related dissertation by the Researcher.
NON-INTERFERENCE

The Instructor agrees to refrain from helping the Participants to perform any of their assigned tasks related to the experiment, will discourage Participants from working together, and will discourage Participants from seeking help from anyone other than the Researcher. All questions about the study should be directed to the Researcher at 383-8166 (work) or 375-4271 (home).

COSTS OF MATERIALS

The Researcher will arrange to pay for all handouts and materials.

INFORMED CONSENT

The Researcher agrees to provide the Participants with an explanation of (a) the nature and purpose of the study in general terms, (b) the procedures to which the Participants will be exposed, (c) the extent of Participant confidentiality and anonymity and (d) the clear statements that a Participant may, without prejudice to him/her, withdraw from the study, cease participation, and/or have his/her data destroyed at any time of his/her choosing.

INCENTIVES FOR THE PARTICIPANTS TO COMPLETE THE STUDY

The Instructor agrees to actively encourage the Participants to complete the study. Examples of the types of phrases the Instructor will use to encourage the Participants to complete the study include: (a) "Research is very important and you are making a very important contribution to this particular research study." (b) "You are helping to answer some very important questions about the practical uses of computers for conducting content analysis." (c) "Participating in research studies helps advance the body of knowledge for a field and leads to the development of new and innovative practices."

The Instructor also agrees to give specific incentives for Participants who complete all four tasks of the experiment (Check those that apply.)

1. Replace one regular assignment. (Check if applicable.)
2. Provide Bonus points; (Check if applicable.)
3. Award the higher of two grades in "borderline" cases. (Check if applicable.)
4. _______
MODIFICATIONS TO THE AGREEMENT

This agreement constitutes a general framework for participation in a content analysis study. Procedures not explicitly described in this agreement shall be conducted according to the Methods section of the dissertation proposal related to this study. Modifications to the explicitly stated sections of this agreement may be made only with the mutual consent of the Instructor and the Researcher. Such modifications shall be documented as a signed and dated amendment to this agreement.

SUMMARY OF BENEFITS TO THE PARTIES

The Instructor/Participants will receive: (a) a classroom lecture on the theories and methods of content analysis; (b) experience with a real-world content analysis problem; (c) experience with a true experimental design research study; and (d) a debriefing about the study after data collection has been completed. The Researcher will receive a portion of the sample of participants needed to conduct a dissertation-related experiment.

SIGNATURES

________________________  _______________  ____________________  _______________
Instructor                     Date                     Researcher                   Date
BIBLIOGRAPHY


Barcus, F. E. (1959). Communications content: Analysis of the research 1900-1958: A content analysis of content analysis. *Dissertation Abstracts*, 20, 3279. (University Microfilm No. 60-143)


Cronbach, L. J. (1963). Course improvement through evaluation. Teachers College Record, 64, 672-683.


