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

by

Richard D. Frisbie

Evaluation Center
Western Michigan University
Kalamazoo, Michigan 49008

Paper presented at the meeting of the American Educational Research Association, San Francisco, April 16, 1986.

THE USE OF MICROCOMPUTERS TO IMPROVE THE RELIABILITY AND VALIDITY OF CONTENT ANALYSIS IN EVALUATION *

Richard D. Frisbie
Western Michigan University

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 paper is intended to promote such an understanding so that evaluation practitioners can better address applied content analysis problems used in their work.

The following discussions are organized into two main sections. First, a general model for conducting an evaluation effort that focuses on information, actions, and standards of quality is presented. This framework highlights key relationships between evaluation and content analysis. It also provides the basis for the experimental design elements of the study. The second section is used to summarize a two-part experimental study that focuses on the reliability and validity of (1) developing a new content analysis category system, and (2) coding responses based on an established category system.

A General Model for Conducting an Evaluation Effort

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

* Paper presented at the meeting of the American Educational Research Association, San Francisco, April 16, 1986.

authors use verb forms of these concepts to represent actions, they are better thought of in their noun forms for this paper--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, (d) present working definitions of evaluation and content analysis, (e) consolidate the above concepts into the general model, and (f) discuss some content analysis tasks related to the model's basic actions that can be implemented with the use of microcomputers.

Key Components of Evaluation and Content Analysis Information

Even though evaluation can be used to develop descriptions and judgments about some object while content analysis can be used to develop descriptions and inferences about some object, 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, it is possible to clarify the conceptual similarities and differences between the various types of information by examining the underlying conceptual contrasts and identifying other terms that more clearly reflect these contrasts. This was accomplished by identifying the applicable underlying contrasts, combining them to clarify their relationships, and then 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.

Many authors (e.g., Gove, 1971, p. 786; Guba & Lincoln, 1981, p. 35; Joint Committee, 1981, p. 12; Morris, 1969, p. 453; Stake, 1967, p. 109; Worthen & Sanders, 1973, p. 19) contend evaluation information contains two fundamental components, often called descriptions and judgments. This surface distinction is based on the underlying philosophical distinction between epistemology and ethics. 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. Thus, the underlying evaluation contrast is between knowledge and value statements.

Content analysis authors often discuss information in terms of descriptions and inferences (e.g., Berelson, 1952, p. 18; Holsti, 1969, p. 14; Osgood, 1959, p. 36; Stone, Dunphy, Smith, & Ogilvie, 1966, p. 5). This surface distinction is based on the means through which information about some object is acquired. Descriptions are based on sensory input of some sort,

commonly 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. Thus, the underlying content analysis contrast is between observations and logic.

The results of reconstructing the components of evaluation and content analysis information that distinguish them from each other in terms of the underlying conceptual relationships involved 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.

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. In addition, the underlying concepts are arranged so that observations and value statements do not overlap. This precludes any components from representing the naturalistic fallacy. [1]

The first level represents the two basic components of evaluation information. For the remainder of this paper 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.

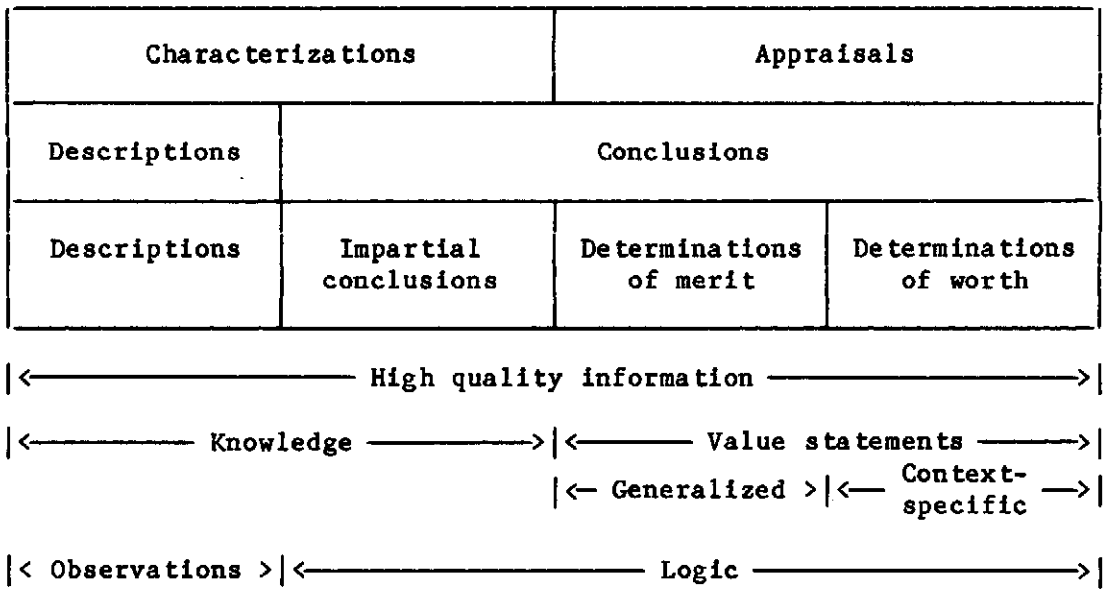


Figure 1

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

The third level represents four subcomponents that are parts of both evaluation and content analysis information. These subcomponents include descriptions, impartial conclusions, determinations of merit, and determinations of worth.[2] 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--examples of the naturalistic fallacy--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.

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 the evaluation literature (Brinkerhoff, Brethower, Hluchyj, & Nowakowski, 1983, p. v; ERS Standards Committee, 1982, p. 11; Joint Committee, 1981, pp. xvii-xx; Stufflebeam et al., 1971, p. 40) and one from the content analysis literature (Krippendorff, 1980, p. 169), were used to identify these actions.

Although the above sources identify several actions that apply to both evaluation and content analysis efforts, 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. Four of the actions focus on processing some kind of information while two focus on the effort itself. These six actions and their main focus are summarized in Table 1.

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

Table 1

Six Basic Actions for Conducting an Evaluation or Content Analysis Effort

Action	Focus
* Delineating	Information
* Obtaining	
* Providing	
* Applying	
* Managing	The effort
* Evaluating	

needed and how it will be acquired. Obtaining information involves the general action of acquiring it in its "raw" state and transforming it to a usable state. Providing information involves the general action of delivering it to the appropriate audiences. Applying information involves the general action of using it for intended or unintended purposes.

Second, the two actions of managing and evaluating an effort are suggested. Managing an effort involves the general action of ensuring all required functions are performed appropriately. Evaluating an effort involves the general action of characterizing and appraising it. Meta-evaluation is another term that can be used for evaluating evaluation efforts.

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.

The two sources of the most comprehensive standards for evaluation quality are the Joint Committee on Standards for Educational Evaluation (1981) and the Evaluation Research Society (ERS Standards Committee, 1982).

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--it 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. In the Functional Table of Contents the standards are displayed according to major tasks of an evaluation. (pp. 1-2)

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. They include: "(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.

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.

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.

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

Evaluation standards are also 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 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.**

The Model

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

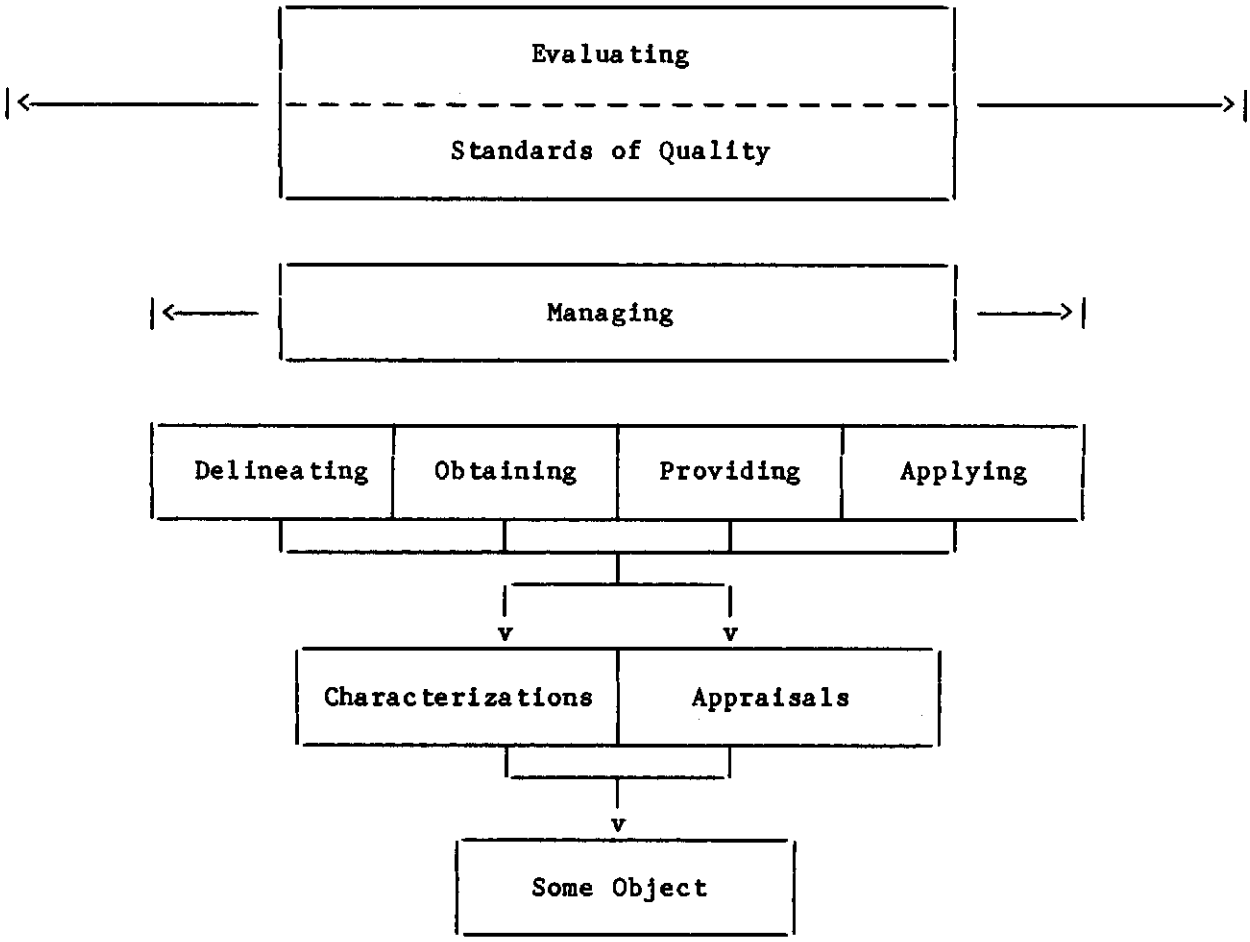


Figure 2

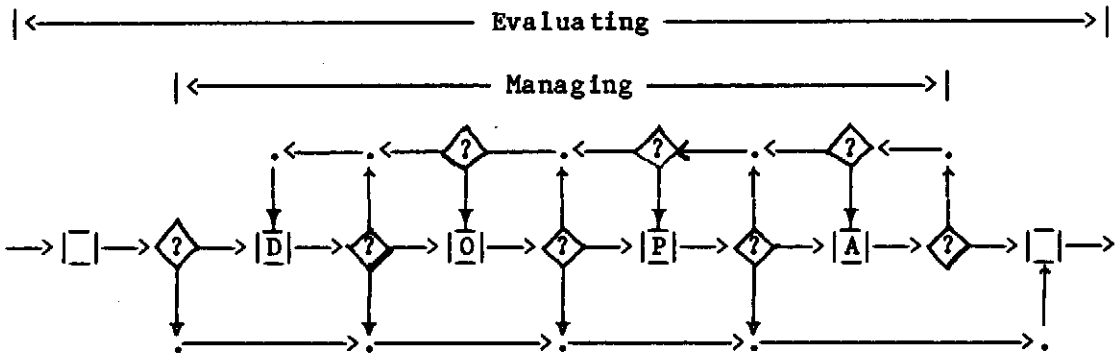
The General Model for Conducting an Evaluation Effort

In this context, evaluation is best thought of as a process that includes six basic actions. Four of them--delineating, obtaining, providing, and applying--focus on processing information. For evaluation, the information includes characterizations and appraisals about some object.

The other two basic actions focus on the effort to process the information. 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 <u>all</u> 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 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.



Legend. |D| = Delineating information, |O| = Obtaining information, |P| = Providing information, |A| = Applying information, |_| = Unspecified action, ◇ = Which action next?

Figure 3

Decision Network Using the General Model for Conducting an Evaluation Effort

This network can also accommodate two ideal types of evaluation information processing patterns that are conceptually incompatible but are probable 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. 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 responsive approach, 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.

Furthermore, the network is hierarchically recursive in nature. That is, 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.

Performing Content Analysis Tasks with Microcomputers

Aside from general project management and support activities like word processing, budgeting, and task planning, computers can be put to three general uses in content analysis efforts (Krippendorff, 1980). These uses 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. 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. It will be discussed in more detail shortly.

Computational content analyses are performed primarily by computer programs rather than by humans. Such programs are simultaneously very complex and overly simplistic. That is to say, 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, Dunphy, Smith, & Ogilvie, 1966).

While Krippendorff and other authors (e.g., Gerbner, Holsti, Krippendorff, Paisley, & Stone, 1969; Holsti, 1969) discuss a number of variations of computer-assisted content analysis, the techniques most useful for survey and discovery can by and large 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 may 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; might be deleted from the list completely.

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 responses to a survey question) 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 occurrences of specified words surrounded by portions of the text in which they occur. This gives the reader an idea of 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 always centered. The line can also be indexed so that 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.

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 passage, 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 usually be sent to a number of destinations, such as the screen, printers, and other data or text 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 in terms of 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.

The Experimental Study

The previous section was used to establish the conceptual relationships between evaluation, content analysis, and microcomputers. This section is used to describe an experimental study in which microcomputers were 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. This section 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 variable, (e) the research hypotheses, (f) the data source, (g) the data analyses employed, (h) the results of the study, and (i) a discussion of the implications of the study.

Review of the Problem

Evaluation practitioners must often collect and analyze responses to a set of spoken or written survey 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 among this set of fixed responses; or, at the other extreme, the questions may be "open-ended," in which the questions are phrased in such a way as 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 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 adequate 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, several content analysis techniques can now be adapted to work with general purpose programs running on relatively inexpensive microcomputers. This means many content analysis 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 makes a difference in the quality of the final results.

The problem then becomes one of determining if available microcomputer programs can be used by evaluation practitioners to help improve the quality of their survey content analysis activities. More formally, the general problem can be stated as follows: How can practitioners use microcomputer programs to improve the reliability and validity of content analyses of responses to open-ended survey questions used in evaluation efforts?

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 survey 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 remedy to this dilemma is to devise 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 survey 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 simulation activity.

The simulation activity placed each study participant in the role of a student research assistant working at a university research center. The director was just called away to an important meeting so 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 five category classification 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. Further details of the simulation are available from the author.

Figure 4 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)

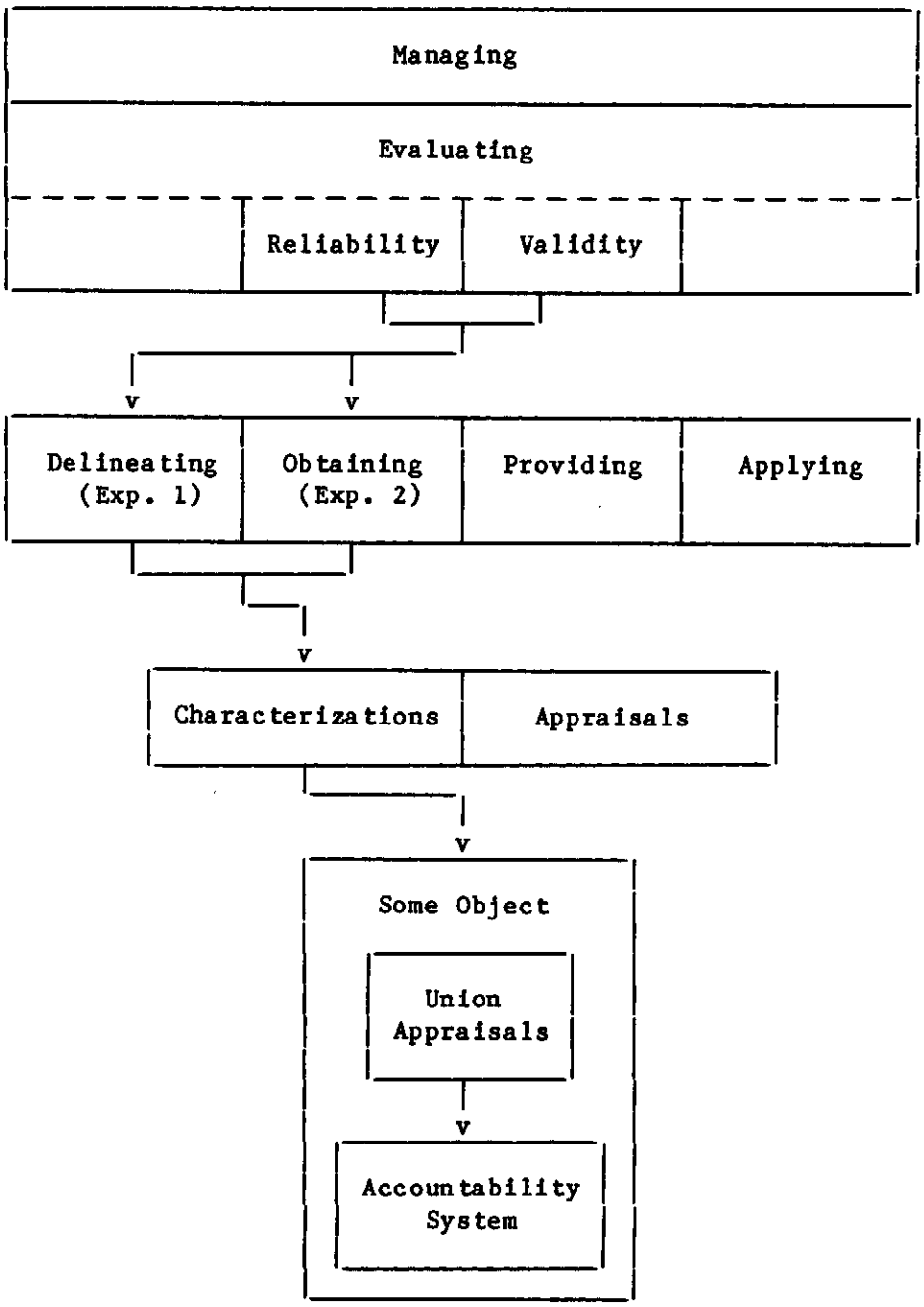


Figure 4

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

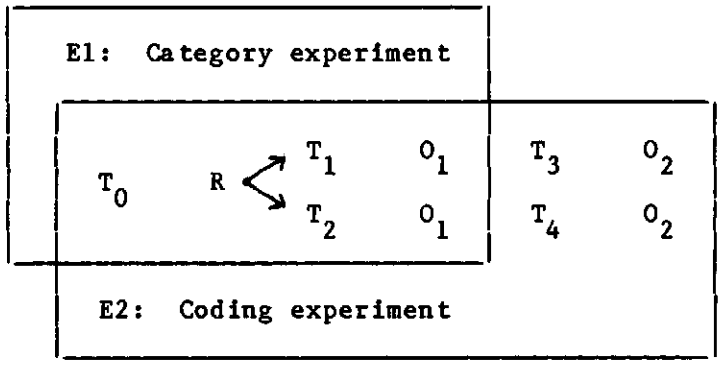
develop a category system, (2) verify the system, (3) code a set of responses, and (4) verify the codes. Both experiments used a posttest only control group design.

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, the experimental group received word counts derived from the responses used in the study while the control group 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 that 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 applicable to 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 5 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. Again, the experimental participants received special assistance coding a set of responses (T_3) while the control participants did not (T_4). At the end of the coding experiment, measures of coding reliability and validity for both groups were recorded (O_2).



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 5

Design of the Study for Two Procedurally Overlapping Experiments

Independent Variable

The independent variable for this study was the possession or lack of possession of outputs from microcomputer programs based on selected content analysis techniques. These outputs included: (a) word counts sorted by frequency of occurrence derived from the responses used in the study (see Appendix A), and (b) the responses sorted and labeled with category identifiers and summaries according to how they were coded by each participant (see Appendix B). Experimental participants received these outputs at specified times during the experiments. Control participants received no word count lists and received response lists in the same order before and after they were coded.

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.

Data Source

The data source for the study was comprised of students enrolled in several College of Education courses at Western Michigan University. These courses represent two educational perspectives. The first perspective is from a research point of view. These courses emphasize research methods like survey research and content analysis. The second perspective is from an educational practitioner's point of view. These courses emphasize skills a teacher would use in elementary and secondary school classrooms. Participants in the study were individually and randomly assigned to either the experimental or control group.

Seventy-four participants completed Experiment 1 and 59 participants completed Experiment 2. These sample sizes met or exceeded the minimum sizes needed to use a pre-determined Type I error level of 0.10, a Type II error level of 0.40 (power of 0.60), and a medium effect size (Cohen, 1977, p. 387).

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 for the experimental and control groups. The validity hypotheses, Hypotheses 2 and 4, were tested by comparing dif ferences between mean total scores for the experimental and control groups.

Measure of Category Reliability

For Hypothesis 1, category reliability is defined as the extent to which the same set of categories are created from the simulation documents and responses used in the category experiment under varying circumstances, at different locations, by different participants. 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. The measure used 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 fifteen Hierarchy categories by the Hierarchy Panel. (Contact the researcher 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. 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.

Measure of Category Validity

For Hypothesis 2, 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. (Contact the researcher for details.) This is analogous to what Krippendorff calls semantical validity (1980, pp. 159-162). 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, in that 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.

Measure of Coding Reliability

For Hypothesis 3, 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 the concept of inter-rater agreement under the measure of category reliability applies here as well.

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.

The measure for coding reliability was derived for each participant with the same Turbo Pascal program use 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, 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.

Measure of Coding Validity

For Hypothesis 4, 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.

Analysis Procedures

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 the use of 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.

Results

The results of the data analyses are summarized in Table 2. The first two hypotheses were based on Experiment 1. They were used to test category reliability and validity. No significant differences between the two groups were found for either of these hypothesis tests. The last two hypotheses were based on the Experiment 2. The third hypothesis test showed a significant difference between the two groups at the 0.001 level. The fourth hypothesis test showed a significant difference at the 0.1 level.

Table 2

Summary of Analysis of Variance Tests for the Four Study Hypotheses

	Hypothesis			
	1. Category reliability	2. Category validity	3. Coding reliability	4. Coding validity
F	1.957	1.479	12.838	3.751
Significance	0.166	0.228	0.001	0.058
Decision	Retain	Retain	Reject	Reject

Thus, it is concluded that output from microcomputer programs 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. It is also concluded that microcomputer output did help

experimental participants more reliably and validly code the open-ended responses in terms of the category system.

Discussion

Based on these results and the relationships between evaluation, content analysis, and microcomputers, it is recommended that microcomputers be used to help practicing evaluators code responses to open-ended survey questions. It is also recommended to conduct further studies that will help determine how microcomputers can be used to more effectively help practicing evaluators create category systems for responses to open-ended survey questions.

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 identifying and organizing a set of key concepts for defining and describing a particular field of study.

The two experiments directly focus on analyzing a set of responses to an open-ended survey question used in a simulated evaluation context. Practicing evaluators can benefit from the positive results of the second experiment in two ways. First, by using computerized content analysis techniques on large bodies of narrative responses to open-ended survey questions, studies that would have been conducted even without the availability of a microcomputer can be conducted more reliably with more meaningful results. Second, new studies can be conducted that otherwise would have been considered too complex or too cumbersome to conduct by traditional, 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 microcomputer technology can be used to enhance the understanding of and practice in each of these fields.

Notes

1. 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--the distinction between merit and worth--observational value statements are not included as an acceptable type of information for this study.

2. Evaluators typically distinguish between two types of value statements (e.g., Guba & Lincoln, 1981; Joint Committee, 1981; Scriven, 1967, 1978) 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 <u>generalized</u> 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 <u>class</u> 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-attribute of university professors. Instead, it should

be taken to mean, generally speaking, university professors with several refereed publications in their field are of value. 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 through the use of 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 information, ethical principles, and logical transformations used to make determinations of merit should be explicitly stated.

The concept of worth represents context-specific value statements. 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 clearly impossible for them to be inherent attributes of an object.

APPENDIX A
WORD COUNT LIST FOR EXPERIMENTAL PARTICIPANTS ONLY

ISU Accountability Study for Hometown Public Schools

WORD COUNT LIST OF WORDS OCCURRING MORE THAN ONCE IN FREQUENCY ORDER
FOR ALL RESPONSES TO THE OPEN-ENDED QUESTION

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 processer. The list contains all of the words that appeared more than once, sorted in order by frequency of occurence. Maybe these lists will give you a few leads to follow when you start to develop the five categories of responses.

#	WORD	#	WORD	#	WORD
108	THE	9	THEY	5	TESTS
72	TO	9	PEER	5	ETC
59	AND	9	THEIR	5	FAR
55	IN	9	OUR	5	USED
53	A	9	OTHER	5	WORK
47	OF	9	TOO	5	AM
44	IS	8	BY	5	S
40	I	8	BEEN	5	LEARNING
36	ARE	8	WAS	5	WANTS
34	ACCOUNTABILITY	8	WORKING	5	ANY
33	TEACHERS	8	ACCOUNTABLE	5	SEEMS
32	NOT	8	STUDENTS	4	RATE
31	BE	8	ADMINISTRATION	4	RATING
28	AS	8	IF	4	HOWEVER
28	IT	7	ONE	4	SHOULD
28	THAT	7	SOME	4	BELIEVE
26	SYSTEM	7	RATINGS	4	SCORES
23	FOR	7	IDEA	4	AT
19	WE	7	PEOPLE	4	WILL
18	WITH	7	WHEN	4	INDIVIDUAL
18	BUT	7	THIS	4	USEFUL
18	HAVE	6	MANNER	4	THOSE
16	GOOD	6	OR	4	CHILDREN
16	ON	6	MAKE	4	TESTING
13	WHO	6	DOES	4	BAD
13	HOMETOWN	6	HE	4	HIGH
12	HAS	6	BECAUSE	4	JUST
12	CAN	6	SCHOOL	4	EDUCATIONAL
12	DO	6	OUT	4	SEE
11	NO	6	WAY	4	AMONG
11	THERE	5	YOU	4	EXPECTED
10	ALL	5	EXCELLENT	4	HAT
10	AN	5	THAN	4	US
10	HUMAN	5	LIKE	4	FROM
10	TEACHER	5	PROGRAM	4	ANOTHER
10	FEEL	5	WERE	4	BEST
10	MANY	5	SO	4	EVEN
10	EACH	5	EDUCATION	4	TEACHING
9	WOULD	5	VARIABLES	4	INTO
9	WHICH	5	MUST	4	STUDENT

#	WORD
4	TAKE
3	TOGETHER
3	SHARING
3	HIS
3	IT'S
3	OBJECTIVES
3	RATED
3	NONE
3	UP
3	THINK
3	SHOW
3	ONLY
3	DONE
3	CONTROL
3	PRINCIPAL
3	PRESSURE
3	THINGS
3	CONCERNED
3	DEALING
3	FORCE
3	CAUSE
3	AGAINST
3	ADMINISTRATORS
3	CHILD
3	BETWEEN
3	GROWTH
3	BEINGS
3	MORE
3	COULD
3	ME
3	MY
3	EVER
3	SUPERINTENDENT
3	WRONG
3	COMPETITION
3	RATHER
3	STAFF
3	CANNOT
3	GROUP
3	HERE
3	LITTLE
3	MODEL
3	USE
3	FEELING
3	HOW
3	BEYOND
3	MUCH
3	BECOME
3	LEVEL
3	GOAT
3	MOST
3	I'M
3	RESULT
3	LOWER
3	VERY
3	IDEAS
3	THEM
2	CHECK
2	ABOUT
2	SELF
2	HIM
2	NEED
2	PRODUCT
2	ACHIEVE
2	SAME
2	EVALUATIONS
2	CONSIDERATION
2	B
2	IMPLEMENTED
2	BUSINESS
2	ATTITUDES
2	BACKGROUNDS
2	ORDER
2	MIGHT
2	SIMPLY
2	SLOW
2	POSITIVE
2	HUMANITY
2	GET
2	POSSIBLE
2	RESPECT
2	DON'T
2	TRYING
2	EVERYONE
2	PROBLEMS
2	MANAGEMENT
2	TRY
2	SHORT
2	THING
2	VALUE
2	IMPORTANT
2	SITUATION
2	MATERIALS
2	EXPERIENCE
2	MAY
2	CLASSROOM
2	FACTORS
2	VIEWED
2	POOR
2	BASED
2	UPON
2	DIDN'T
2	ITS
2	COMPONENTS
2	WHAT
2	HELD
2	FIGHTING
2	GONE
2	DECENCY
2	STANDARDIZED
2	DEMEANING
2	PROBLEM
2	DURING
2	AFRAID
2	PROGRESS
2	DOESN'T
2	RESENT
2	JOB
2	JOKE
2	INCREASING
2	CLASSES
2	NEVER
2	JUDGING
2	TOOL
2	EVALUATION
2	LOST
2	ROOM
2	WELL
2	TIME
2	YEAR'S
2	CARE
2	CLASS
2	WHOLE
2	UNDER
2	HOME
2	SINCE
2	PLAY
2	LOW
2	MERIT
2	ECONOMIC
2	PAY
2	BETTER
2	SHARE
2	THEMSELVES
2	PART
2	MEETING
2	RESPONSIBLE
2	OTHERS
2	PRESENTLY

APPENDIX B

RESPONSES SORTED BY CATEGORY FOR EXPERIMENTAL PARTICIPANTS ONLY

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 ubnpleasantness 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 opperssive and stifling.

#	Category Old	New	Response
45	A	__	Accountability is important, but not as a fear developing tool. I was among the four who resigned.
47	A	__	The accountability system is a good idea gone bad.
50	A	__	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.

5	B	__	Competition is increasing for high scores on HAT tests and good ratings by principals.
13	B	__	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.
15	B	__	Teachers stay in room and do not share.
19	B	__	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.
23	B	__	It seems to imply that we must be in competition with our colleagues in order to be good teachers.
46	B	__	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.
49	B	__	Accountability here is backbiting, and dividing (as he wants) teachers.

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

6	C	__	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*******."
8	C	__	There are too many variables that enter in to make it work.
9	C	__	There are too many variables in the educational system for accountability to work.
18	C	__	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 diffrent 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)

References

Andrén, G. (1981). Reliability and content analysis. In K. E. Rosengren (Ed.), Advances in content analysis (pp. 43-67) (Vol. 9, Sage Annual Review of Communication Research). Beverly Hills: Sage.

Berelson, B. (1952). Content analysis in communications research. New York: Free Press.

Borland International. (1983). Turbo Pascal reference manual version 2.0 [Computer program manual]. Scotts Valley CA: Author.

Brinkerhoff, R. O., Brethower, D. M., Hluchyj, T., & Nowakowski, J. R. (1983). Program evaluation: A practitioner's guide for trainers and educators (Vol 1: Sourcebook/Casebook) Boston: Kluwer-Nijhoff.

Budd, R. W., Thorp, R. K., & Donohew, L. (1967). Content analysis of communications. New York: Macmillan.

Carney, T. F. (1972). Content analysis: A technique for systematic inference from communications. Winnipeg, Canada: University of Manitoba Press.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). New York: Academic Press.

ERS Standards Committee. (1982, September). Evaluation Research Society standards for program evaluation. In P. H. Rossi (Ed.), Standards for evaluation practice (pp. 7-19). (New Directions for Program Evaluation No. 15). San Francisco: Jossey-Bass.

Digital Equipment Corporation. (1984a). Word processing using list processing [Computer program manual]. Maynard, MA: Author.

Digital Equipment Corporation. (1984b). Word processing using sort [Computer program manual]. Maynard, MA: Author.

Gerbner, G., Holsti, R., Krippendorff, K., Paisley, W. J., & Stone, P. J. (Eds.). (1969). The analysis of communication content: Developments in scientific theories and computer techniques. New York: John Wiley.

Gove, P. B. (Ed.). (1971). Webster's third new international dictionary of the English language (unabridged). Springfield, MA: Meriam.

Guba, E. G., & Lincoln, Y. S. (1981). Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches. San Francisco: Jossey-Bass.

Harrison, J. (1967). Ethical naturalism. In P. Edwards (Ed.), The encyclopedia of philosophy (Vol. 3, pp. 69-71). New York: Macmillan.

Holder, W. (1982). The word plus: Spelling checker with automatic correction [Computer program manual]. San Diago: Oasis Systems.

Holsti, O. R. (1969). Content analysis for the social sciences and humanities. Reading, MA: Addison-Wesley.

Janis, I. L. (1965). The problem of validating content analysis. In H. D. Lasswell, N. Leites, & Associates (Eds.), Language of politics: Studies in quantitative semantics. Cambridge: MIT Press.

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

Kaplan, A., & Goldsen, J. M. (1965). The reliability of content analysis. In H. D. Lasswell, N. Leites, & Associates (Eds.), Language of politics: Studies in quantitative semantics. Cambridge: MIT Press.

Krippendorff, K. (1980). Content analysis: An introduction to its methodology. Beverly Hills: Sage.

Morris, W. (Ed.). (1969). The American Heritage dictionary of the English language. Boston: American Heritage Publishing and Houghton Mifflin.

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

Osgood, C. E. (1959). The representation model and relevant research methods. In I. de Sola Pool (Ed.), Trends in content analysis. (pp. 33-88). Urbana: University of Illinois Press.

Patton, M. Q. (1980). Qualitative evaluation methods. Beverly Hills: Sage.

Patton, M. Q., French, B., & Perrone, V. (1976, August). Does accountability count without teacher support? An assessment of the Kalamazoo Public Schools accountability system from the perspective of teachers. Minneapolis: University of Minnesota, Minnesota Center for Social Research.

Ratliff, W. (1982). dBASE II: Assembly language relational database management system [Computer program manual]. Culver City, CA: Ashton-Tate.

Scriven, M. S. (1967). The methodology of evaluation. In R. Tyler, R. Gagne, & M. Scriven, Perspectives of Curriculum Evaluation (pp. 39-81), (AERA Monograph Series on Curriculum Evaluation No. 1). Chicago: Rand McNally.

Scriven, M. S. (1978). Merit vs. value. Evaluation News, 1(8), 20-29.

Stake, R. E. (1967). The countenance of educational evaluation. Teachers College Record, 68, 523-540.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). The general inquirer: A computer approach to content analysis. Cambridge: MIT Press.

Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Merriman, H. O., & Provus, M. M. (1971). Educational evaluation and decision-making. Itasca, IL: Peacock.

Tyler, R. (1949). Basic principles of curriculum and instruction. Chicago: University of Chicago Press.

U.S. Department of Education. (1981). Application of grants under handicapped personnel preparation program. Washington, D.C.: Author.

U.S. General Accounting Office. (1978). Assessing social program impact evaluations: A checklist approach. Washington, D.C.: Author.

Worthen, B. R., & Sanders, J. R. (1973). Educational evaluation: Theory and practice. Belmont, CA: Wadsworth.