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

(p. 333). 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).