CHAPTER IV

RESULTS

Introduction

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

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

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

Effects of Microcomputer Output on Category Development

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

The independent variable was the possession or lack of possession of two types of microcomputer output: (1) a word count list sorted by frequency of occurrence derived from the complete set of responses used in the simulation, and (2) half of the responses sorted according to how they were coded by each participant, and labeled with participant-developed category identifiers and summaries. Forty-two randomly-assigned experimental participants received word count lists at the beginning of the first task while control

participants never received such lists. Experimental participants received responses sorted and labeled according to their own categories at the beginning of the second task. Control participants (32) received unsorted and unlabeled responses at that time.

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

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

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

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

Category Reliability

Category reliability was defined as the extent to which the same set of categories was created from the simulation documents and responses used in the category experiment under varying circumstances at different locations, by different participants. The median proportion of agreement was used as the measure of category reliability

for each participant because: (a) it reflects the extent to which each participant agreed with the rest of his or her treatment group about which five categories should be included in the response classification system, and (b) analysis of variance can be used to test the null hypothesis related to group mean scores on category reliability when it is operationalized as a proportion of agreement.

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

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

The null hypothesis relevant to research hypothesis 1, category reliability, is as follows: no difference exists between the mean of

the median proportion of category agreement scores for the participants who received specialized microcomputer output and the mean of such scores for the participants who received no specialized microcomputer output. A one-way analysis of variance was used to test this null hypothesis.

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

Table 9

Group Means and Standard Deviations for Hypothesis 1, Category Reliability

Group	N	Observed mean	SD
Experimental	42	0.47	0.16
Control	32	0.42	0.16

Table 10

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

	Critical comparison				
Group	Observed	Chance	t	Critical value	Decision
Experimental	0.47	0.053	16.89	2.02	Reject
Control	0.42	0.053	12.88	2.04	Reject

Table 11

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

Source	Sum of squares	Degrees of freedom	Mean square	F	Sig. of F
Between	0.050	1	0.050	1.957	0.166
Within	1.854	72	0.026		
Total	1.905	73			

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

Category Validity

Category validity was defined as the extent to which the set of categories created by participants in the category experiment agreed with the set of categories created by the Response Panel. The total number of agreements with the Response Panel was selected as the measure of category validity for each participant.

The null hypothesis relevant to research hypothesis 2, category validity, is as follows: no difference exists between the mean total number of category agreements with the Response Panel for the participants who received specialized microcomputer output and the mean of such scores for the participants who received no specialized microcomputer output. A one-way analysis of variance was used to test this null hypothesis.

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

Table 12

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

Source	Sum of squares	Degrees of freedom	Mean square	F	Sig. of F
Between	2.474	1	2.474	1.479	0.228
Within	120.405	72	1.672		
Total	122.878	73			

Table 13

Group Means and Standard Deviations for Hypothesis 2, Category Validity

Group	N	Observed mean	SD
Experimental	42	3.12	1.21
Control	32	2.75	1.39

As summarized in Table 12, no statistically significant difference ($p \leq 0.10$) in mean category validity scores was found between those groups of participants who did or did not receive specialized microcomputer outputs. Therefore, the null hypothesis was retained. Thus, the difference between the obtained mean score for the experimental group (3.12) and the obtained mean score for the control group (2.75) is considered to have occurred by chance alone.

Effects of Microcomputer Output on Response Coding

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

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

For Task 3, experimental participants (30) received the 50 original responses sorted and labeled according to their own categories, as updated during the previous task, plus 50 new responses at the end of the list. For Task 4, the experimental participants

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

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

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

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

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

Coding Reliability

Coding reliability was defined as the extent to which the same codes were assigned to the coding experiment responses under varying circumstances, at different locations, by different participants.

The median proportion of agreement was used as the measure of coding reliability for each participant because: (a) it reflects the extent to which each participant agreed with his or her treatment group about how the 100 responses should be coded, and (b) analysis of variance can be used to test the null hypothesis related to group mean scores when coding reliability is operationalized as a proportion.

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

The null hypothesis relevant to research hypothesis 3, coding reliability, is as follows: no difference exists between the mean of the median proportion of coding agreement scores for the participants who received specialized microcomputer output and the mean of such scores for the participants who received no specialized microcomputer

output. A one-way analysis of variance was used to test this null hypothesis.

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

Table 14

Group Means and Standard Deviations for Hypothesis 3, Coding Reliability

Group	N	Observed mean	SD
Experimental	30	0.82	0.05
Control	29	0.75	0.09

Table 15

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

	Critical comparison				
Group	Observed	Chance	t	Critical value	Decision
Experimental	0.82	0.20	67.92	2.04	Reject
Control	0.75	0.20	32.91	2.05	Reject

As shown in Table 14, the observed mean score for the experimental group is 0.82. The observed mean score for the control group is 0.75. As indicated in Table 15, both of these observed scores are higher than would be expected by chance alone ($p \leq 0.05$). Thus, the

corresponding null hypotheses were rejected. Finally, the analysis of variance summarized in Table 16 indicates the difference between the two observed mean scores is greater than chance ($p \leq 0.001$). As a result, the null hypothesis stating no difference exists between the experimental and control group means on category reliability was rejected. Thus, the experimental proportion of agreement (0.82) is higher than the control proportion of agreement (0.75).

Table 16

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

Source	Sum of squares	Degrees of freedom	Mean square	F	Sig. of F
Between	0.063	1	0.063	12.838	0.001 **
Within	0.280	57	0.005		
Total	0.343	58			

** Significant at 0.001 level

Coding Validity

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

The null hypothesis relevant to research hypothesis 4, coding validity, is as follows: no difference exists between the mean total number of coding agreements with the Response Panel for the

participants who received specialized microcomputer output and the mean of such scores for the participants who received no specialized microcomputer output. A one-way analysis of variance was used to test this null hypothesis.

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

Table 17

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

Source	Sum of squares	Degrees of freedom	Mean square	F	Sig. of F
Between	307.205	1	307.205	3.751	0.058 *
Within	4668.829	57	81.909		
Total	4976.034	58			

* Significant at 0.10 level

Table 18

Group Means and Standard Deviations for Hypothesis 4, Coding Validity

Group	N	Observed mean	SD
Experimental	30	86.63	7.46
Control	29	82.07	10.44

As summarized in Table 17, the observed difference between mean coding validity scores for the group who did and the group who did not receive specialized microcomputer output <u>is</u> statistically

significant ($p \leq 0.10$). Therefore, the null hypothesis was rejected. As noted in Table 18, the obtained mean score for the experimental group (86.63) is higher than the obtained mean score for the control group (82.07). This difference is considered to be greater than what would have occurred by chance alone, if the null hypothesis is true.

Comparisons of Those Who Did or Did Not Complete Each Experiment

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

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

Table 19

Characteristics of Students Enrolled in the Six Study Classes

Characteristic	All students		Completed experiment 1				Completed experiment 2			
			Yes		No		Yes		No	
	N	%	N	%	N	%	N	%	N	%
Total number	125	100.0	74	59.2	51	40.8	59	47.2	66	52.8
Treatment group										
Experimental	65	52.0	42	33.6	23	18.4	30	24.0	35	28.0
Control	60	48.0	32	25.6	28	22.4	29	23.2	31	24.8
Class										
ED 322a	30	24.0	13	10.4	17	13.6	8	6.4	22	17.6
ED 322b	28	22.4	17	13.6	11	8.8	10	8.0	18	14.4
ED 450/455	13	10.4	12	9.6	1	0.8	11	8.8	2	1.6
ED 516	24	19.2	12	9.6	12	9.6	10	8.0	14	11.2
ED 602	14	11.2	6	4.8	8	6.4	6	4.8	8	6.4
EDLD 663	16	12.8	14	11.2	2	1.6	14	11.2	2	1.6
Employed										
Full time	49	39.2	30	24.0	19	15.2	23	18.4	26	20.8
Part time	45	36.0	29	23.2	16	12.8	23	18.4	22	17.6
Not employed	28	22.4	15	12.0	13	10.4	13	10.4	15	12.0
Missing	3	2.4								
Degree program										
Undergraduate	74	59.2	42	33.6	32	25.6	29	23.2	45	36.0
Graduate	48	38.2	32	25.6	16	12.8	30	24.0	18	14.4
Missing	3	2.4								
College of Major										
Arts & sciences	40	32.0	22	17.6	18	14.4	17	13.6	23	18.4
Bus/Engr/Fine	13	10.4	6	4.8	7	5.6	4	3.2	9	7.2
Education	67	53.6	44	35.2	23	18.4	36	28.8	31	24.8
Missing	5	4.0								

Table 19 (continued)

Characteristic	All students		Completed experiment 1				Completed experiment 2			
			Yes		No		Yes		No	
	N	%	N	%	N	%	N	%	N	%
Expected grade										
A	64	51.2	44	35.2	20	16.0	35	28.0	29	23.2
BA	22	17.6	12	9.6	10	8.0	9	7.2	13	10.4
B	22	17.6	11	8.8	11	8.8	9	7.2	13	10.4
Missing	17	13.6								
Proportion of total hours completed										
0.00 to 0.40	37	29.6	24	19.2	13	10.4	19	15.2	18	14.4
0.41 to 0.80	34	27.2	15	12.0	19	15.2	12	9.6	22	17.6
0.81 to 1.00	15	12.0	11	8.8	4	3.2	7	5.6	8	6.4
1.01 to 1.40+	5	4.0	3	2.4	2	1.6	2	1.6	3	2.4
Missing	34	27.2								
Proportion of full time enrollment										
0.00 to 0.40	47	37.6	33	26.4	14	11.2	25	20.0	22	17.6
0.41 to 0.80	24	19.2	14	11.2	10	8.0	11	8.8	13	10.4
0.81 to 1.00	31	24.8	15	12.0	16	12.8	12	9.6	19	15.2
1.01 to 1.40+	19	15.2	12	9.6	7	5.6	11	8.8	8	6.4
Missing	4	3.2								
Content analysis studies										
0	117	93.6	70	56.0	47	37.6	55	44.0	62	49.6
1	5	4.0	4	3.2	1	0.8	4	3.2	1	0.8
Missing	3	2.4								
Other studies										
0	114	91.2	69	55.2	45	36.0	55	44.0	59	47.2
1 to 4	8	6.4	5	4.0	3	2.4	4	3.2	4	3.2
Missing	3	2.4								
Total studies										
0	111	88.8	67	53.6	44	35.2	53	42.4	58	46.4
1 to 4	11	8.8	7	5.6	4	3.2	6	4.8	5	4.0
Missing	3	2.4								

students, (b) treatment group membership, (c) enrolled class, (d) employment status, (e) current degree program, (f) college of major, (g) expected grade for the class, (h) the proportion of the total hours completed for the applicable degree program, (i) the proportion of full time enrollment for the degree program, (j) the number of content analysis studies in which the student participated, (k) the number of other studies in which the student participated, and (l) the total number of studies in which the student participated.

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

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

Table 20

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

	Variable				
	Treatment	Class	Employment	Degree	College
Chi-square	1.210	16.747	0.866	0.819	2.353
Degrees of freedom	1	5	2	1	2
Significance	0.271	0.005	0.649	0.366	0.308
Decision	Retain	Reject	Retain	Retain	Retain

Table 21

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

	Variable					
	Expected grade	Proportion hrs. completed	Proportion FTE hours	Number C.A. studies	Number other studies	Total number studies
F	1.365	0.530	0.962	0.810	0.183	0.447
Sig.	0.246	0.469	0.330	0.370	0.670	0.505
Decision	Retain	Retain	Retain	Retain	Retain	Retain

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

Table 22

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

	Variable				
	Treatment	Class	Employment	Degree	College
Chi-square	0.004	24.688	0.218	5.436	2.903
Degrees of freedom	1	5	2	1	2
Significance	0.949	0.000	0.897	0.020	0.234
Decision	Retain	Reject	Retain	Reject	Retain

Table 23

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

	Variable					
	Expected grade	Proportion hrs. completed	Proportion FTE hours	Number C.A. studies	Number other studies	Total number studies
F	0.142	0.514	0.005	2.091	0.166	0.676
Sig.	0.707	0.476	0.944	0.151	0.685	0.413
Decision	Retain	Retain	Retain	Retain	Retain	Retain

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

Summary

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

Table 24

Summary of Analysis of Variance Tests for the Four Study Hypotheses

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

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

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