To the Graduate Council:

I am submitting herewith a dissertation written by Tracy L. Cross entitled "An Analysis of the Testing Practices of College Professors." I have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Education.

Schuyler W. Huck, Major Professor

We have read this dissertation and recommend its acceptance:

Accepted for the Council:

Vice Provost
and Dean of the Graduate School

AN ANALYSIS OF THE TESTING PRACTICES OF COLLEGE PROFESSORS

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Tracy L. Cross

December 1988

Copyright © Tracy L. Cross, 1988
All rights reserved

Dedicated to
my mother, the late Eva Ingle Cross
and my wife, Jennifer Riedl Cross

ACKNOWLEDGEMENTS

I wish to give special thanks to Dr. Schuyler W. Huck who directed this dissertation and served as a mentor throughout my doctoral program. Sky instilled in me confidence that this project could and should be conducted. I also wish to thank my dissertation committee: Dr. Laurence J. Coleman, who provided inspiration to complete the project along with practical advice on how to do so; Dr. Howard Pollio, who provided the expertise in phenomenology and a general "world view" which made this project stronger; and Dr. E. Ohmer Milton, who was a vast source of knowledge (particularly concerning the Wormley Hotel) and a kindred spirit.

To Lee Humphreys and Barbara Wickersham, I wish to express thanks for their financial support and use of the facilities at the Learning Research Center of the University of Tennessee including the staff needed to complete many of the arduous tasks in the project: Linda Gipson, who organized the effort, Laurel Humphreys, Beth Donaghy, Pam Littlejohn, Susan Lentz, Rena George, and Karen Flatt.

I wish to thank Kevin Richter, Jeff Risinger, and Eric Dixon for many hours of hard work which culminated in the final reduction of the telephone interview data.

My thanks to Dr. Edward Roeske who once told me as a fledgling graduate student that I had the potential to obtain a Ph. D.

and consequently allowed me to start realizing this dream. Thanks also to John Habel, who provided often needed prompts to finish this project.

My appreciation to Jennifer Cross, who provided immeasurable support during this project and throughout the last 13 years of my life. And finally, to Ian and Keenan Cross who provided the ultimate inspiration to finish.

ABSTRACT

The purpose of this study was to investigate and analyze college professors' testing practices and beliefs about different types of tests. In addition, the researcher sought to determine and make explicit the processes and the variables that college professors consider when deciding what types of tests to use in the classes they teach.

To develop a description of the college classroom testing phenomenon, a questionnaire was sent to a representative sample of college professors from across the United States. The questionnaire asked the faculty to submit information about the types of tests used in the smallest and largest classes taught during the last completed academic term. To attain an understanding of how college professors decide what types of tests to use in the classes they teach, telephone interviews were conducted with selected questionnaire respondents. Personal interviews with three questionnaire respondents were conducted in order to provide descriptions of the experience of testing from a phenomenological perspective.

Discrepancies were found between the types of tests many professors believe to be the most effective and the types they actually use. Logistical concerns on the part of the professors weighed heavily in their decisions about what tests to use. Class size appears to be the most salient variable considered by professors as

they design the evaluation component of a course. Multiple choice tests were used most often, while essay tests were the tests of choice. In addition, it was found that there are four common themes to the experience of testing college students: an awareness of others, limitations of time, the test's fit with the class material, and the relationship of tests with grades.

TABLE OF CONTENTS

CHAPTER	PAGE
1. INTRODUCTION	1
The Testing of College Students	1
Research on the Effects of Testing	3
Research on Test Practices	5
Significance of the Study	5
Methodology: A Brief Description	6
Basic Assumptions	8
Summary	10
2. REVIEW OF THE LITERATURE	11
The Evolution of Classroom Testing	11
How Different Types of Tests Affect Students' Behaviors and Learning	25
The Current Testing Practices of College Professors	32
3. METHODS AND PROCEDURES	40
Selection of Subjects	41
Survey Instrument	45
Categorization of Survey Responses	49
Telephone Interviews	51
Lazzara's Interview-Analysis Technique	52
Conducting the Telephone Interviews	60
The Personal Interviews	62
The Pilot Study: Multiple Choice Tests	67
Summary	68
4. RESULTS	70
The Survey Respondents	70
The Survey Responses	75

CHAPTER PAGE

4. (Continued)

The Paradox Groups. 82
Results of the Telephone Interviews. 90
Results of the Personal Interviews. 101
The Awareness of Others in Testing. 105
The Experience of Time in the Testing Process. 113
Fit with Course Material and Objectives. 115
The Experience of Grades in the Testing Process. 119
Use of Machine-Scorable Answer Sheets. 123

5. DISCUSSION. 126

Test Types Used in the College Classroom. 127
Is Multiple Choice Test Usage on the Increase?. 128
Do Professors Use the Test Type They Consider Best?. 131
Professors' Decisions to Use a Test Type. 137
The Experience of Testing Students. 139
Implications of the Study. 140
Suggestions for Future Research. 144
Limitations. 145

LIST OF REFERENCES. 147

APPENDIXES. 155

A. The Survey Instrument. 156
B. Letter to Subjects Accompanying Survey 157
C. Letter to Testing Service Directors 158
D. Letter to Software Companies. 159
E. Telephone Interview. 160
F. Full Arguments of Telephone Interviews 164
G. Personal Interview. 184
H. Second Reduction of the Personal Interviews 193

VITA. 202

LIST OF TABLES

TABLE		PAGE
1.	Test Types Used at a Large University	33
2.	Test Types Used at a Small University	34
3.	Extent to Which University of Tennessee Faculty Use Test Types	36
4.	Number of Survey Respondents by School Size	72
5.	Number of Survey Respondents by Academic Area	73
6.	Number of Survey Respondents by Teaching Level	74
7.	Rankings of Test Types Used by Respondents	80
8.	Rankings of Test Types on Which Grades are Based	81
9.	Ranking of Paradox Groups by Teaching Level	84
10.	Ranking of Paradox Groups by Academic Area	86
11.	Ranking of Paradox Groups by School Size	87
12.	Paradox Categories	89
13.	Telephone Interviews in Total Agreement Paradox Group	96
14.	Telephone Interviews in Total Disagreement Paradox Group	97

TABLE PAGE

15. Telephone Interviews in Large Class Agreement Paradox Group 98

16. Telephone Interviews in Small Class Agreement Paradox Group 99

17. Telephone Interviews in Class Agreement Paradox Group 100

18. Significant Statements in the Personal Interviews. . 102

19. Number of Machine-Scorable Answer Sheets at Five Large Schools 124

CHAPTER 1

INTRODUCTION

Evidence that testing was used as a means of assessing students' knowledge and learning has been found as early as 2200 B.C (Dubois, 1966). Four thousand years later, testing still plays an important role in the lives of students. While some research has been conducted on the effects of tests on students' behavior and learning, little research exists on the testing practices of college professors. This study was an attempt to develop an accurate picture of the testing phenomenon as it exists in the college classroom.

This chapter describes the significance and purpose of the study, its limitations and basic assumptions. The chapters that follow provide an in-depth description of related research, the structure of this study, and the presentation and interpretation of the collected data.

The Testing of College Students

Each year approximately twelve million American college students are administered tests in the courses they take. Almost all of the tests are developed by the course instructor. Assuming each student takes an average of three courses per quarter, and three

tests are required in each course, slightly over one hundred million tests are administered to college students in an academic year (Milton, Pollio, & Eison, 1987). This estimation represents only tests used to provide students' grades, it does not take into account diagnostic tests or any other types. Given the vast number of tests that college students are exposed to, it is obvious that testing plays a major role in most students' college education.

Since the turn of the century, various types of tests have been developed for use in the classroom. Prior to their development, professors had at their disposal only a few types (essay, oral, short answer and derivations of each). These were often very cumbersome to use since they took a great deal of time to score. They have also been criticized for their "subjective" nature. With the birth of the field of educational measurement, new types were developed which offered remedies to the time problems associated with the older tests, and which were considered to be more "objective" in their ability to measure student learning and/or achievement.

These new testing methods have proliferated within our education system from preschool programs to professional and graduate schools. This proliferation has spawned questions about the efficacy of these tests. Since the early nineteen hundreds, a small number of research projects have been conducted to determine the effects that tests have on college students.

Research on the Effects of Testing

One of the earliest research projects about the effects of testing was undertaken in 1934. Terry found that students study differently for different types of tests (1934). Several studies which followed Terry's study sought to determine if different types of tests administered to students affected the manner of preparation the students engaged in prior to the exam and consequently affected the type and degree of student learning. It was found that different types of tests do prompt different study techniques prior to the exams and, consequently, result in different types and degrees of learning (Douglass & Talmadge, 1934; Meyer, 1934; Meyer, 1935; Meyer, 1936). After reviewing the available published research in 1964, Balch stated that: "Understanding both how and why testing influences learning is incomplete, but the use of an evaluating instrument and the nature of the instrument used influences students' learning."

In addition to determining that test type used affects students' preparation, along with the degree of learning and their ability to maintain what they have learned, some research supports the assertion that individual test questions vary in the level of cognitive abilities demanded of the students (Bloom, 1956). In the classic text Taxonomy of Educational Objectives, Bloom and his associates (1956) describe six levels at which test questions can be classified: Knowledge, Comprehension, Application, Analysis, Synthesis, and

Evaluation. This taxonomy has been used in several research studies for evaluating the quality of test questions. Writing test questions at differing levels of Bloom's Taxonomy is common advice to builders of tests. Doing so, however, is very difficult for experienced test constructors and even more so for the novice. Complicating the process of writing good test questions is that the level at which a test question is written is contingent upon the prior knowledge the student may have in the area being tested (Scannell and Tracy, 1975; Bloom, 1956). For example, a question written at the "Application" level could be lowered to the "Knowledge" level if the person being tested prepares for the question and merely has to recite what he/she memorized before the test.

Making matters even more difficult is the fact that developing the various types of test questions--essay, multiple choice, true-false--at the levels of Bloom's Taxonomy have inherent varying degrees of difficulty. For example, writing essay questions at the evaluation level of Bloom's Taxonomy is believed to be far easier than constructing evaluation questions using the true-false format. Some test construction experts believe that it is possible to construct questions at all levels of Bloom's Taxonomy using any type of test question; they also note the difficulty and question the utility of doing so (Frederiksen, 1984; Gronlund,1985). Even with research on the effects of tests on learning, it appears that college professors still

develop questions without an appreciation of the consequences of their testing practices (Frederiksen, 1984).

Research on Test Practices

While there have been some studies conducted to determine the influence of testing on learning, little research exists on the testing practices of college professors. The studies that have been published suggest that the types of tests developed since World War II tend to be those most often used in the college classroom (Spencer & Stallings, 1970; McNames, 1977; Committee on Testing, 1978; Lunneborg, 1977; Trani, 1979; Chase & Wakefield, 1984). Two of these studies limited their samples to faculty at a single university and, consequently, should not be considered conclusive. With so little research available on the testing practices of college professors, it is impossible to have an accurate perspective of the overall testing phenomenon.

Significance of the Study

In the past, little research has been conducted with the purpose of describing college professors' testing behavior. While the research done on the impact of testing in the classroom is important, it can only be put into its proper perspective if we also know how college professors use tests. Without knowing what is being

practiced in college classrooms, it cannot be determined to what degree college students are being affected by tests.

As evidenced by past research on the effects of various types of tests on learning, in-class testing could be considered the driving force that greatly influences the quality of education students receive in college. Despite the existence of this research, few professors receive formal training in pedagogy. Conceivably, we can then assume that many of the one hundred million tests developed by college professors are subject to flaws having profound implications for the learning process. Without training in proper test construction, professors may opt for those that are most conveniently developed and graded rather than those most appropriate for evaluating and influencing student learning. The lack of research allows college professors to chart their own course in the way in which they test their students. Their choices are based on factors other than psychometric theory and include such factors as their own experience with testing as students, expediency, departmental policy, etc.. By developing an accurate picture of the college classroom testing phenomenon, the effects of testing and the ways tests are applied can be put into perspective.

Methodology: A Brief Description

The purpose of this study was to investigate and analyze college professors' testing practices and beliefs about different types

of tests. The study also sought to determine and make explicit the processes and variables that college professors consider when deciding what types of tests to use in class.

By answering the following questions, this research is designed to explicate the current testing practices of college professors in order to achieve an accurate perspective of the overall testing phenomenon. Specifically, the present study sought to answer the following questions:

1. What types of tests are used by college professors?

2. Is the use of multiple choice tests on the increase?

3. Do college professors use the type(s) of tests they believe to be the best method(s) for assessing students' learning in the classes they teach?

4. If some college professors do not use the type(s) of tests that they believe to be the best for assessing students' knowledge, why not?

5. How do college professors decide what types of tests to use in the classes they teach?

6. How do college professors describe their experience of the testing process?

To develop a description of the college classroom testing phenomenon, this study began with the construction of a questionnaire to be sent to a representative sample of college professors across the United States. The questionnaire asked subjects to submit information about the types of tests used in the smallest and largest classes taught during the last completed academic term. Faculty from large, medium and small colleges and universities comprised the questionnaire sample. To attain an understanding of how professors decide what types of tests to use in their classes, telephone interviews were conducted with selected respondents. In-depth interviews with three respondents were conducted to provide a description of the experience of testing from the professors' perspective. The information gathered from the questionnaires, telephone and personal interviews provide evidence necessary to describe college classroom testing.

Basic Assumptions

When conducting any study, a researcher must rely on assumptions that make the project feasible. In this case there were three major assumptions upon which the study was based. The first was that subjects would accurately describe their testing practices

on the survey instrument. In an attempt to ground subjects' responses in prior patterns of behavior as opposed to requesting that they describe their thoughts about testing in the abstract (removed from actual experience), the survey specifically asked them to provide information about their testing practices during the last completed academic term. All returned surveys with information which was not based on recent experience, incomplete, or illegible were eliminated from the sample. During telephone interviews, attempts were made to validate information received on the surveys.

The second assumption was that the subjects selected for the telephone interviews could articulate how they made the decision about what type(s) of tests to use in the two classes described on the survey instrument. Even though the information being sought is tacit in nature, it was assumed that subjects could accurately identify reasons behind their decisions. The interviewing method used during this phase of the data gathering procedures was designed to gather such tacit information.

The third assumption was that the sample of college professors selected to be subjects in this study was representative of the population of college professors across the United States. Subjects were selected using a stratification process which was likely to produce a representative sample. A detailed description of the selection process can be found in the third chapter.

Summary

Tests are a large part of the education process in today's college classrooms. Research has shown that the type of test used in a class can have a profound effect on the learning accomplished by the students, yet there has been little research done on the types of tests that college professors use. In fact, there is not an accurate picture available to describe how college professors feel about the different test types and how they determine what test type they will use in a class. In consideration of the substantial effect that their choice can have, this is a very significant piece of information.

This study was designed to provide a complete description of the college classroom testing phenomenon. This introduction has presented the facts which support the need for such a study. Now that the stage has been set, the next chapter provides detail on related research followed by a description of the methodology of the study. The fourth chapter contains the results obtained. Using the information received, a concluding description of the college classroom testing phenomenon is set forth in the final chapter.

CHAPTER 2

REVIEW OF THE LITERATURE

The research found relevant to this study provides historical perspective on the subject of testing in the college classroom. By presenting a description of factors influencing the evolution of classroom testing, it becomes evident how the current state of affairs in testing is a logical extension of past events. Due to the progression of events, there has been some research conducted on the effects of testing on students. Following the historical presentation, a summary of the research that has been conducted on the ways in which different types of tests affect students' behavior and learning will be described. To conclude the literature review, those studies regarding the current testing practices of college professors will be discussed.

The Evolution of Classroom Testing

DuBois (1966) gives credit to the Chinese for developing the first formalized testing procedures about 2200 B.C.. According to Smallwood (1935), Martin describes the process as "... shut [ting the student] up for a day and a night, each in his narrow cell, they produce each a poem and one or two essays on themes assigned by the chancellor" (p. 62). The Chinese methods for testing remained

relatively unchanged from that time until the early 1900's. In contrast, American schools spent much more time involving students in rote learning activites which tended to be verbal. In addition, the American schools saw major changes in their teaching and testing methods due to the technological advancements of the times. One of the earliest technological advancements that seemed to have a significant effect on American schools' testing procedures was the advent of the blackboard. During the 1700's and early 1800's, schools in the United States did not have blackboards. This device, which enabled group instruction to take place, came into existence in 1809 (Parker, 1912). Cureton (1971) cites the blackboard as an important element in the change of educational practices. When discussing the time prior to the advent of the blackboard, she states:

> ...until well into the nineteenth century, instruction and examining were mostly oral. Either the class responded in concert, or the pupil came forward to the teacher's desk and recited alone.

The second major change in testing practices in American schools came as a consequence of the common usage of the metal pen. Before 1828 writing was taught only to develop penmanship, due largely to the difficulty of keeping the goose quills sharpened (Thayer, 1928). Because of the availability of the new metal pens, written tests became common during the 1850's. The tests were

either essays written about particular themes assigned by the teacher or were written recitations.

Because of advances in available educational materials, educators could concentrate less on trying to get their messages across to students and more on methods of assessing their learning. During the late 1700's, the educational leaders of schools in the United States became interested in developing a procedure for ranking students. In 1783 Yale University became the first college to implement a formalized procedure for ranking students. Other colleges and universities soon followed by implementing their own grading procedures. Harvard, Mount Holyoke, William and Mary, The University of Michigan and The University of Georgia had their own methods of assigning grades established by the early 1800's (Smallwood, 1935). By the 1820's more than half of the academic institutions in the United States were using the five-point grading scale (Nicolson, 1917).

The task of assigning grades to students is an important factor in the evolution of different methods of testing students. New types of tests needed to be developed to provide the degree of certainty in the evaluation of student performance that was demanded by the times. No longer would students be given the nod to go on to higher grades based on teacher opinion. Educators were in the early stages of calling for more precise measures of student performance and for new methods of finding that precision. This desire for precision was an early indicator of the shift away from using tests to evaluate

learning and towards using them for the purpose of ranking students.

In the mid- to late-1800's, educational leaders like Horace Mann presented their opinions about testing to the public. Many of these leaders cited the virtues of written examinations over oral exams. According to Caldwell and Coulter, Mann stated that written exams prevented the "officious interference of the teacher" (1923). These educational leaders instigated the idea that educational testing requires a scientific professionalism which to date had not been considered.

In 1845 prior to Horace Mann's stint as the Secretary of the Massachusetts Board of Education, a Boston school committee prepared questions from textbooks used in the schools of the day. The questions were to be used to assess the schools' effectiveness and to assess student learning. The committee described the purpose and philosophy of the test as follows:

> It was our wish to have as fair an examination as possible; to give the same advantage to all to prevent leading questions; to carry away, not loose notes, or vague remembrances of the examination but positive information, in black and white; to ascertain with certainty what scholars know, as well as what they did not know; to test their readiness at expressing their ideas upon paper; to have the positive and undeniable evidence of their ability or inability . . . (p. 26).

This committee's testing philosophy was an early indication that educators wanted the tests taken by students to be ones which asked questions with answers that were "in black and white," right or wrong, not left to the judgment of the teacher. This attitude gave rise to what would later be known as objective testing. Objective tests appeared to provide the precision the committee desired for the evaluation of students' performance since the "subjective" judgment of the teacher was eliminated. An example of the types of questions the committee designed for the test is:

> Do the waters of Lake Erie run into Lake Ontario, or the waters of Lake Ontario into Lake Erie? (p. 53).

In addition to this type of question, which by today's standards would be described as a type of multiple choice question, the test also asked questions to which the student provided factual information (Cureton, 1971). This committee's test is very important because it provides one of the earliest examples of tests which required written responses but did not require written essays. This new form of test was designed to measure students' ability to memorize facts. This test question would fall into the lowest level of Bloom's Taxonomy of Educational Objectives called "knowledge" (Bloom's Taxonomy will be discussed later in this chapter). The "level" of information required by these test questions is important because it became a template for later tests used throughout the

country. Tests used by teachers today often have questions which look quite similar to the one described above.

The overall mean score for pupils taking the committee's test was 35.5 percent correct. When talking about the results of the test, Horace Mann wrote that: "The facts disclosed by the Reports, have spread through the city a general and deep feeling of sorrow and mortification..." (p. 50). An implication of this remark is that this new type of test had been able to uncover the poor performance of students that had been suspected by the committee, but to date had been masked by inept assessment techniques used by teachers: namely, written essays. By the 1890's, this new method of testing had become common in the assessment of school effectiveness in the Northeastern United States.

Although they had been for years the preferred method for testing students, written exams began receiving criticism after the development of the Boston school committee's test. In 1912 and 1913 educational researchers Starch and Elliot published three articles which cited the unreliability of teachers' grading on written exams which took the form of essays. They described ranges of forty to fifty points in grades on the same papers in English, arithmetic, and history. They concluded that essay papers could not be scored accurately beyond five point intervals. Other studies which soon followed concluded that five or ten grades were about as many as could be distinguished (Rugg, 1915). Because of these studies, many of the remaining colleges and universities which were using a 100

point scale for student grading were influenced to convert to letter grades. The studies also added to the growing concerns over the utility of essay examinations.

During the early twentieth century, growing interest in statistics and in measurement also prompted the development of new types of tests which could be assessed using these new comparative techniques. These tests "forever changed the methods of examinations" (Cureton, 1971, p. 4). A group of psychologists headed by Robert Yerkes, president of the American Psychological Association, developed the Army Alpha and Beta tests which were used to test the ability of thousands of people who had been inducted into the Army as soldiers. Vast numbers of soldiers were tested and assigned to duties on the basis of their performance on the tests.

The Army tests were very efficient in the time needed to take them, to score them, and the ease with which one could administer them to hundreds of test takers at once. This ease of administration represents the progress that had been made in the process of testing students since the 1700's when one student would be called before the teacher to recite his or her lessons. The psychologists who developed these tests were able to show, using statistics, that their tests were both valid and reliable. The format of the Army tests was multiple choice. The term "objective" became synonymous with this new form of test since the test could be scored by numerous graders with the final grades being identical. Multiple choice tests were

believed to be superior to essay tests since they were not fraught with the problems "inherent" in essay tests--their inability to be reliably graded. Many people generalized the fact that multiple choice tests could be graded reliably to mean that they were objective tests in the truest sense--without subjectivity. Those in the measurement profession generally agree that since test questions are written by instructors, they are necessarily subjective when developed.

In 1919 the two psychologists who developed the Army Alpha and Beta Tests, Robert Yerkes and Lewis Terman, developed a standardized national intelligence test for students (Chapman, 1980). Rather than being exclusively a military tool, the standardized test was now being used in the education system. The climate of the times seemed to be right for these new types of tests since they allowed efficient testing and subsequent ranking of large numbers of students. The desire for such new tests was partly due to the growth in the public schools. From 1910 to 1920 the enrollment in elementary schools had jumped from 19.3 million students to 23.3, while enrollment in secondary schools had risen from .9 to 2.2 million (U.S. Bureau of Census, 1975, p. 368). Since many of these new students were immigrants, it was believed by many educators that a new, efficient, scientific method for testing students was greatly needed because many of the immigrants did not speak English well and some had little to no formal education.

For years the developers of the Army tests, as well as others

developing new statistical techniques, received a great deal of media attention. In addition, in 1909 a book entitled Laggards in Our Schools appeared which declared that a great many students who were attending school were too old for their grade and that the schools were to blame for their inefficiency in discriminating among students' performance. This book, along with Starch and Elliot's works on the unreliability of grading, was the primary reason educational leaders of the day called for "more objective and reliable measurement" as a means to improve the efficiency of the schools (Haney, 1984).

Various standardized tests were developed during the 1920's and 1930's. In 1933 Buros published a book entitled Educational, Psychological and Personality Tests of 1933 and 1934. The book only had 44 pages. By 1938 the revised edition, known as the Mental Measurements Yearbook, had 400 pages describing over 4000 tests (Buros, 1938). Haney describes the 1930's as the decade of the "proliferation of educational testing" (1984, p. 609).

Although educational tests and more sophisticated statistical procedures were continually being developed, the Great Depression seemed to cause issues in testing that had once been debated with great passion to lay dormant (Gould, 1981). The Second World War brought testing back into the lives of the layperson since once again there was a great need for tests to help decide the various roles incoming soldiers would play in the Army. However, there was not

the same fervor as before since mental testing had clearly become well established in education in the United States (Haney, 1984).

Since testing had become such a common component in the lives of students in the United States, issues within testing were often left to test experts to decide. Sophisticated statistical equations often provided the answer to criticisms levied at the tests. More and more tests were developed by organizations such as the Educational Testing Service (ETS) which came into existence in 1947.

In 1965 the U.S. Congress passed a bill called the Elementary and Secondary Education Act (ESEA). The ESEA was an "important watershed in the history of American education" (Bailey and Mosher, 1968, p. vii). The ESEA called for federal monies to be spent on the basis of program evaluations that called for norm-referenced tests (NRTs) to be used to discern the success of the educational programs. At the same time a controversy had arisen among educational testers as to the utility of the NRTs. Criterion-referenced tests (CRTs) were becoming popular and it was believed by many measurement experts that the NRTs deviated too far from the academic skills and knowledge that should be measured (Madaus, et al., 1979).

Popham was one of the earliest supporter of CRTs. He published numerous articles that criticized the NRTs and suggested that CRTs are the proper tests to use with students (Popham, 1978a; Popham, 1978b; Popham & Husek, 1969). Popham (1978b) listed three major shortcomings of NRTs:

1. Because of the way they are constructed, NRTs can result in unrecognized mismatches between what is being tested and what is being taught.

2. Because of their generality, NRTs provide imprecise instructional targets.

3. The techniques used in building NRTs, which help to insure considerable response variance among examinees, can lead to the elimination of items covering important topics or skills just because all examinees get the items right or wrong.

During the 1970's the discontent with NRTs grew. Numerous articles and books criticizing NRTs, intelligence tests, and standardized tests in general were published. The articles collectively published in a book entitled The Myth of Measurability (Houts, 1977) levied "a devastating attack on standardized testing" (Bollenbacher, 1976). Ralph Nader's consumer advocate group, along with other collectives, including the NAACP, began to call for "truth in testing" laws to be passed which would protect the rights of students taking the standardized tests. They demanded that ETS reveal to the general public examples of questions that have been used on the tests as well as providing test takers' scored copies of their exams with all the correct responses indicated. In 1979 New York became the first state to pass "truth in testing" legislation (Haney, 1984). The laws uncovered that "at least in minor measure the reformers' argument [concerning unfair practices against students] was correct" (Haney, 1984). In 1980 Nader's group issued

a report entitled The Reign of ETS: The Corporation that Makes Up Minds (Nairn and Associates, 1980).

At the same time that ETS was being ridiculed by proponents of civil liberties, national achievement and aptitude test scores were declining. Scores on the most popular test used for admittance to college, the Scholastic Aptitude Test (SAT), had fallen steadily over a fifteen-year period. Great concern was voiced by test administrators, professional educators, and concerned citizens in the country's newspapers as well as in the professional journals. In addition to the falling SAT scores, grades for college students rose. This inconsistency seemed to represent serious flaws in teachers' grading systems--the standardized tests were indicating something much different from the grades teacher assigned. The only way to ensure that students were performing adequately was to verify their performance with more standardized testing. Much like the movement that took place in the mid-1800's, those concerned began calling for minimum competency testing of students. Several states passed laws between 1976 and 1980 which required students to pass a test proving that they had at least a minimum competency in academic subjects as a prerequisite to receiving a high school diploma (Haney, 1984; Cross & Huck, in press).

One of the most important events that affected the testing practices in the United States began in 1979 when a group of ten black high school students sued the state of Florida over the state's newly implemented minimum competency requirements. The ten

students had been denied high school diplomas as a consequence of failing the graduation test. The case, Debra P. v. Turlington, lasted several weeks with many leaders in the field of testing arguing against one another on issues dealing with whether or not the state of Florida had carried out the necessary procedures to build a valid test as prescribed by the standards manuals developed by those same experts (Haney, 1984). The court found in favor of the plaintiffs but neither side was satisfied with the subrulings. The case was appealed in 1981 to the U.S. Court of Circuit Appeals. The judge ruled that the state of Florida had failed to prove that the minimum competency test had a close match with the curriculum of the Florida schools and that students must be trained to take these multiple choice tests if states plan to give them in this fashion. These findings sent shock waves through high schools and colleges across the United States. This type of litigation was seen by some to force standardized testing on schools, teaching for those tests, and a movement even further away from essay-type testing. It was clear that the testing practices of schools, individual teachers, and college admissions' officers were going to be scrutinized for some time to come. One effect of this type of litigation may be to make naive educators and laypersons believe that the format that most standardized tests use (multiple choice) is the preferred, most appropriate method to evaluate students.

Summary

Testing practices have undergone major changes from the first known tests by the Chinese some 4000 years ago to those of the present day. Effected strongly by improvements in the educational materials available, public opinion and massive increases in the number of students being tested, these testing practices have shifted over the years from oral or written essay tests to tests which are believed to be more objective. Objective tests have been supported by the government, and their results are easily interpreted by the public partially because they are supported by statistical estimates of reliability and validity.

Even though there is widespread support, some organizations have voiced opinions against objective testing. A move away from such tests has not been seen yet, and objective testing is the "test of choice" for many present-day educators. An education in today's society no longer has meaning without the verification of a standardized test. A diploma is not proof that a student has performed well; a passing grade on a standardized test is. This attitude has affected not only academics; professionals, too, must pass tests to become licensed. It makes sense that many college professors would be seduced into believing that objective tests are the most appropriate for evaluating students' learning since virtually all competency or licensure tests take this form.

How Different Types of Tests Affect Students' Behaviors and Learning

During the early part of the twentieth century, the United States saw an unprecedented attempt by educators to develop new types of tests that would be used to assess students' educational abilities, academic achievement, vocational interests, and "innate" intelligence. Along with the development of these new tests, a few educators became interested in how the tests used in the classroom to assess student learning affect the students taking them. The research conducted supports the idea that students are affected by the tests they are given. Not only do students prepare differently for each of the different test types, their performance on a test is affected by their preparation. In 1964 Balch evaluated seventy articles and studies on the effects that an evaluating instrument has on students' learning. In conclusion, he stated: "Understanding both how and why testing influences learning is incomplete, but the use of an evaluating instrument and the nature of the instrument used influence students' learning."

One of the earliest published studies concerned with the influence of the new test types which were rapidly infiltrating the American classroom was carried out by Terry at The University of Alabama in 1934. Terry gave 135 students in his psychology class a checklist of study method procedures and asked them to indicate how they prepared for the different types of tests. Information was

gathered about multiple choice, true-false, completion, and essay tests. From student responses Terry concluded that students study differently for different types of tests.

Not long after Terry's study was published, two University of Minnesota researchers, Douglass and Talmadge, published an article entitled "How university students prepare for new types of examinations." Douglass and Talmadge (1934) also found that college students reported that they study differently for different types of tests. For tests which require the recognition of a correct response, students focused their attention on the exact wording of the readings; while for tests which require them to recall information, the students reported that they looked for relationships and trends in the readings.

In the mid-1930's Meyer at the University of Michigan conducted detailed studies to determine how students studied for tests and whether their study methods affected test performance. Students were divided into four groups, each of which was asked to prepare for a particular type of test: i.e., completion, essay, multiple choice, and true-false. The testing was over two mimeographed chapters about the Civil War with which the students were not familiar.

Meyer and the researchers working with him perused the notes taken by the students while they prepared for the different types of tests. He found that the students who studied for completion, multiple choice, and true-false tests did not attempt to

gain an overview of the material. Instead, they concentrated on trying to memorize facts and particular statements. He also found in reviewing the notes that six study methods were revealed: 1)the underlining of words, phrases, and sentences in the booklets; 2)the listings of names, places, dates, and numbers; 3)the taking of random notes; 4)the making of summaries in paragraph form; 5)the drawing of maps; and 6)the framing of practice test questions. Meyer (1935) claimed that the most significant finding of his analysis of notes was that a smaller percentage of students who had prepared to take the essay test used underlining while a greater percentage of them made summaries and prepared maps than did the students in any other group. This suggests that essay tests encourage students to go to greater lengths in their efforts to learn the material than the other objective test types.

The students in Meyer's study actually took all four of the different test types. Five weeks later, the students took the four types of tests again. Meyer found that the students who prepared for either the essay or the completion tests scored better on all four forms of the exam at the initial testing and again five weeks later when the testing was repeated (1935).

In 1936 Meyer conducted a second study at the University of Michigan. This study was similar to the first study, but three features were added: the students heard the material they were to study and be tested over discussed in class, a study period was built into the course, and the students received four sample test questions

prior to taking the tests. One sample of each of these test types was given to the students: completion, essay, multiple choice, and true-false.

Once again, it was found that the students prepared for the tests in different ways using the same techniques described in the previous study (Meyer, 1936). The focus of this study was, however, on student performance. Meyer was trying to determine if preparing for a particular type of test better prepared students as measured by their performance on the test. He measured their performance immediately after they prepared for the test and again one month later. In this study, Meyer found that the students who prepared for the essay exam scored better on all four types of exams than did the students who prepared for any of the other three types of exams. These results held true when all students were retested four weeks later.

More than thirty years after Meyer's studies, similar research was conducted in England by Thomas and Augstein (1970). Sixty students were divided into four groups of fifteen. As in the previous studies, students were told to prepare for particular types of tests: completion, multiple choice, and a summary paper of the information. Each of the students studied information about the genetic code and later took all three types of tests. The researchers found that the students who prepared for the summary paper performed better on all three types of tests than did the other

students. One week later, the students were retested with the same results.

In 1975 Felker and Dapra conducted a study at the University of Pittsburgh in which they had ninety-five undergraduates study fifteen segments of a textbook. The students were asked to answer one of two types of questions during the study process. Once again, the researchers found that those students who answered comprehension or problem-solving multiple choice questions scored higher on the problem-solving test than the students who answered simple verbatim, recall multiple choice questions.

While the previous studies support the notion that the type of test students prepare for will affect their performance on any test type, Halpin and Halpin (1982) set out to determine whether having a test of any kind versus having no test in a class would change students' study behaviors. They told some students that they were going to be given a test while others were told "study in order to learn." Halpin and Halpin found that the students who prepared for a test scored significantly higher than the students who were told to study to learn. Six weeks later, all students were retested. The test results were the same as in the initial testing. The conclusion drawn from this study is that tests of any kind affect student learning.

In 1937, Wert conducted a study in a zoology class. Wert developed several equivalent forms of a test which measured achievement in five skill areas: 1)naming animal structures, 2)stating functions of structures, 3)knowing terminology and other

facts, 4)interpreting new experiments, and 5)applying principles to new situations. All students took the tests at the end of the course, one year later and two years later. Wert found that there were significant losses in performance on the test over the two year period for the first three skill areas. Conversely, performance in the fourth (interpreting new experiments) and fifth (applying principles to new situations) skill areas remained high. According to the results of Wert's study, testing for information requiring higher levels of comprehension encourages greater retention of the information. Testing for only lower levels of information does not encourage the students to retain the material.

An important gap in educational testing research was filled by Benjamin Bloom's work of the mid-1950's. His book, entitled Taxonomy of Educational Objectives, described qualitative differences in types, or levels, of learning (Bloom, 1956). The levels Bloom developed are often used as the metric to evaluate the level at which the different types of test questions are written. The ability to gauge the skills needed to answer test questions has proven to be an invaluable asset to test developers. As a consequence of Bloom's work, it has been demonstrated that many objective test questions require students to operate at the lowest levels of the Taxonomy. When considering Wert's study on students' ability to retain material, this finding has grave implications for our education system.

In 1984 Norman Frederiksen, a well-known researcher with

the Educational Testing Service, published an article entitled "The Real Test Bias" in which he summarized the research of numerous scholars. The evidence presented supported his claim that tests affect both teacher and student performance. He proclaimed that the real bias in educational testing was the tendency for teachers to test in a very narrow vein of what needed to be taught and consequently had a tendency to teach to the test. Another component of the bias in testing was that the teacher-made multiple choice tests given to students at all educational levels tend not to measure the more complex cognitive abilities. Therefore, only certain skills (ability to recall specific information) are taught to students.

Summary

The proliferation of objective tests in the classroom stimulated much research on the effects that these tests have on the students who take them. In the studies conducted, it was found that students prepare differently for the different test types. Based on students' performance on the tests, it seems that the most effective studying is done when preparing for essay or similar test types. Retention of information seemed to be improved among students preparing for an essay type of exam as well.

This body of research indicates that students benefit more from the preparation for and taking of the test types previously

criticized for their subjectivity. Even so, educators continue to prefer tests considered more objective. As indicated by the previous research, this preference for objectivity and the level at which test questions are written may retard the quality of learning in our educational systems.

The Current Testing Practices of College Professors

Since the testing of college students has been an ongoing phenomenon for hundreds of years, it stands to reason that college professors would have developed a detailed description of the testing practices of their profession. Unfortunately, this is not the case. Few studies have been published which describe the testing practices of college professors. At least two possible explanations could account for this apparent lack of research: (1) little interest exists on this particular topic, or (2) research has been carried out within particular colleges and universities but simply has not been published. It seems likely that the latter explanation is more plausible since two studies carried out within particular schools have been located and are discussed in this section.

Most of the research conducted on this topic has been carried out since the mid-1960's. Two studies with corroborating findings were carried out in the mid-1970's. Both of the studies sought to determine how often professors use the different test types. In 1970 at the University of Illinois, Spencer and Stallings had 1700

college faculty answer surveys which asked them to indicate the frequency of use of particular types of tests: multiple choice, true-false, matching, essay, short answer, and fill-in-the-blank. Spencer and Stallings found that short answer essay, essay and multiple choice were the most often chosen test types. The findings of their study are in Table 1.

Table 1

Test Types Used at a Large University

Recognition Tests	%	Supply Tests	%
multiple choice	14	short answer essay	24
true-false	9	essay	17
matching	7	fill-in-blanks	12

In 1977 McNames conducted a similar study in Kansas. The study surveyed the faculty of two small liberal arts colleges on the frequency with which they used particular types of tests. McNames found that multiple choice, short answer essay and essay were the most often selected test types. McNames' findings are described in Table 2.

The percentages in each of these categories are quite similar to the results of the previous study conducted at a large institution by Spencer and Stallings. This suggests that school size may not have an important role in the decision-making process when professors select the types of tests to administer to students.

Table 2

Test Types Used at a Small University

Recognition Tests	%	Supply Tests	%
multiple choice	17	short answer essay	16
true-false	13	essay	13
matching	9	fill-in-blanks	10

In 1979 a study using a wider sample than the previous two was conducted by Trani. He studied five colleges and universities ranging from a small liberal arts college with an enrollment of 900 students to a large university with an enrollment of 22,000 students. Trani asked faculty members at the schools to indicate how often they used particular types of tests. After combining the subjects' responses of "frequently" and "almost always" on the scale he provided, Trani found that short answer essay tests received the highest percentage: 45.4%. Multiple choice tests were second with 37.5 %. Essay tests were used 25.2% of the time and fill-in-blank items were used 16.1%.

Even though these percentages are similar to previously described research findings, the findings on another part of the same study suggested that the professors may not have answered candidly. Trani asked students at the same schools about their testing experiences. He found that 47% claimed that they had either "never" or "rarely" been exposed to essay exams. This appears to be in direct contradiction to the professors' claims to be using essay tests 25.2% of the time.

In 1978 a study was conducted by a select committee on testing at The University of Tennessee in which questionnaires were distributed to 473 faculty who teach introductory courses. The questionnaire requested information about the professors' testing practices. One hundred and fifty-six returned questionnaires were usable for the study. The professors were asked to what extent they

used particular types of tests in their classes. Table 3 contains the percentages accrued.

Table 3
Extent to Which University of Tennessee Faculty
Use Test Types

Test Type	Rank	Percent
multiple choice	1	17.4
short answer	2	15.9
problem solving	3	15.6
true-false	4	12.8
completion	5	12.6
essay	6	9.6
other (projects, themes)	7	9.3
matching	8	6.8

The responses to this question indicate that all the test types are used to some extent with the most frequently used type being multiple choice and the second most popular test type being short answer. Using a mixture of different item types was also a popular response.

When asked about their classroom practices, approximately 90% of the professors provide students with a test schedule. Approximately 80% provide students with an assignment schedule. Fifty-two percent of the faculty provide students with sample test

questions, while over 90% tell students about item types. Therefore, students tend to know when they will be taking exams and what type of items will be on the test when they finally type them. This knowledge is important since previously reported research has described the effect that preparation for different test types has on the degree and type of learning.

Lunneborg conducted a study in 1977 in which she asked 700 professors at the University of Washington to divide 100 points into the factors they consider when grading their students. Even though the nature of this research question seems rather unclear and unlikely to yield reasonable findings, Lunneborg found that multiple choice test questions had the highest mean percentage rating across faculty members. And as already noted, Lunneborg's findings are quite similar to the findings of other researchers on the same topic.

In 1983 Chase and Wakefield carried out a research project at Indiana University to study faculty evaluation procedures. Chase and Wakefield (1984) found that the tests used most often by faculty to assess students were 1)objective tests, 2)essay tests and research papers (tied), 3)in-class participation. Objective tests in this study were made up of the following item types: multiple choice, true-false, fill-in-blanks, and matching. This study again suggests the common usage of objectively scored tests in the college classroom.

With the limited research suggesting that objective tests are commonly used in college classrooms, a few researchers set out to

ascertain the quality of the tests. In 1968 a study was conducted in which researchers analyzed tests from physics departments in the United Kingdom. Black (1968) found that 15 percent of the questions required working at the application level and 50 percent asked for direct recall of information.

A second study was conducted in 1976 by Semb and Spencer. In this study, seventeen professors from various disciplines within liberal arts at a large university were asked to rate their test questions. The subjects rated 31 percent of their questions as thought provoking. The researcher's analysis showed that only 9 percent of the questions required more than simple recall of information. Semb and Spencer claim that the professors were unaware that they were testing at such a low level of comprehension.

A third study took a different path. The researchers were interested in the quality of multiple choice test questions developed by text publishers since more and more text books were accompanied by test manuals which provide instructors with hundreds of questions to choose from. In the study, several hundred of nearly 2000 available questions were classified by two judges. Both judges, working independently, found that 95 percent of the questions required recognition of simple material (Evans, Dodson, & Bailey, 1981).

Summary

The studies described in this section illustrate the current understanding of the testing practices of college professors to be somewhat amorphous. Researchers in the area suggest that at least three characteristics of this phenomenon have been uncovered: 1)college professors use a wide variety of testing procedures and 2)multiple choice tests tend to be the preferred test type when assigning grades, and 3)the multiple choice tests used typically measure simple recognition of information.

The findings of studies concerning professors' test practices are inconclusive. The sample sizes are not large enough to accurately represent the test practices of professors across the United States. In addition, no attempt was made in these studies to determine why the professors choose the type of tests they do. Due to the small number of studies it is also not clear to what extent college professors use the different types of tests, if professors in different disciplines test differently, and what professors believe about the different types of tests.

CHAPTER 3

METHODS AND PROCEDURES

The research design of this study included a three step data-gathering process: 1)data gathered using a survey instrument, 2)data gathered through telephone interviews, and 3)data gathered through personal interviews. Eleven hundred randomly selected college professors were mailed a package which included the survey instrument and a letter written by the researcher which requested their cooperation in the study. The returned surveys were separated into five categories based on the subjects' responses. From each of the five categories of responses, four subjects were randomly selected to be participants in the telephone interview phase.

After telephone interviews were completed, the researcher randomly selected one professor from each of three levels of schools who had returned one of the surveys, but who had not been a telephone interviewee, to be a participant in the third phase of data-gathering procedures, the personal interviews. The researcher traveled to the three schools to conduct the interviews in person. The interviews focused on the subjects' testing practices, their beliefs about testing, and their experience of the testing process.

In this chapter, a description of the data-gathering procedures is given. In addition, a description of the methods used to select

subjects for the study, an explication of the survey instrument used, and a brief description of the data analysis techniques are discussed.

Selection of Subjects

With all of the different types of tests available to college professors, some subject areas do not afford professors the same latitude of choice in the tests they can administer. For example, it seems reasonable to assume that professors of mathematics would choose problem-solving as the major or exclusive test type. Since one objective of this study was to determine how college professors decide which type(s) of tests to use in the classes they teach, it was important for subjects to possess an ability to choose from a variety of test types. A pilot study was conducted in an effort to hone the survey instrument and to identify those groups of faculty with the widest range of choice when selecting tests for their courses. Only those with considerable latitude were selected as subjects.

In a preliminary study, data were gathered from faculty representing several academic areas. Specifically, the survey was mailed to fifty professors at three universities across the state of Tennessee. The schools used in this pilot study were selected because they matched up well with the three sizes of schools needed for the study; they also were geographically representative of different parts of the state.

Twenty-five surveys were mailed to Memphis State University

faculty, which represented the "Large School" category. East Tennessee State University (ETSU) was designated a "Medium School." Fifteen ETSU faculty were sent surveys. Maryville College represented the "Small School" category and ten faculty members were sent surveys. Approximately sixty percent of the surveys were returned from each school. After surveys were collected, five telephone interviews were conducted with respondents. These subjects were asked to comment on the efficacy of the instrument, suggest improvements to the survey, and to offer a general explanation of their responses.

Following an analysis of the data collected in the pilot study, it was determined that a heavy emphasis should be placed on selecting professors in biology, education, and psychology departments, with a lesser emphasis on other departments. In addition, one of the original six survey questions was eliminated and phrasing in two of the remaining five questions was modified to reduce the degree of ambiguity.

The subjects in the major study were college professors from across the United States who taught in 1987 at accredited post-secondary institutions offering baccalaureate degrees or both baccalaureate and graduate degrees. To select a representative sample of schools, the researcher, using Barron's Profiles of American Colleges, divided all accredited post-secondary institutions located in the United States into three categories based upon their overall enrollment. The categories were: 1) schools with an

enrollment of less than 5000 students, labeled "small schools," 2)schools with an enrollment between 5000 and 15000 students, labeled "medium schools," and 3)schools with an enrollment greater than 15000 students, labeled "large schools."

A total of 31 schools was selected for the study; 10 small schools, 11 medium schools, and 10 large schools. The researcher selected these schools using the table of random numbers located in Ferguson's text Statistical Analysis in Psychology Education. In an attempt to select a representative sample of the overall number of faculty teaching at the three levels of schools, a faculty selection ratio of 5 large school faculty for 3 medium school faculty for 2 small school faculty was generally employed. Using the faculty directory from each school, all faculty who taught in the academic areas of Biology, Education, and Psychology were selected to be subjects in the study. After subjects had been selected from the original three academic areas, a count was made to see how closely the selection process had come to the selection ratio of 5L:3M:2S. To fill the remaining slots, the researcher randomly selected faculty from all other academic areas to obtain a wider representation of academic areas for the study, and to reach the desired ratio. In one instance, however, while selecting subjects from the medium schools, the selection ratio was not close enough to the number of faculty drawn to make filling the slots left with the professors from other areas reasonable. Therefore, an eleventh school was selected to help complete the selection ratio in a reasonable way.

Through the selection process, 558 faculty from large schools, 327 faculty from medium schools and 215 faculty from small schools, for a total of 1100 faculty, were identified as potential subjects. Each individual was mailed a survey and a letter requesting participation in the study.

Anonymity of survey respondents was maintained throughout the study, however, it was necessary to be able to identify respondents in order to analyze the data adequately. In addition, telephone interviews of the selected respondents were to be a significant part of the study. To have this capability, a code was developed and affixed to each survey. As professors were selected for the study, they were assigned a sequential number. This number and the letter L, M, or S, for Large, Medium, or Small, was placed inconspicuously on the survey before it was mailed to the participant. When the completed survey was received, a more complete code was put on the card: Sequential faculty number, L,M, or S for school size, sequential school identifier, B, E, P, or O for academic area (Biology, Education, Psychology, or Other), and a 1,2,3,4, or 5 for faculty level (1 = Full Professor, 2 = Associate Professor, 3 = Assistant Professor, 4 = Other, 5 = Unknown). A sample code would be 509L31P1. The data analysis was dependent on this code for identification of the subjects.

Survey Instrument

The purpose of the survey instrument was to gather information to answer the questions posed by this research project. The types of tests currently being used by college professors and professors' opinions about which test types are best to use for assessing students' learning were the primary concerns of the survey. The researcher suspected that there was an inconsistency (paradox) between what college professors believed about various types of tests and what tests they actually used in their classes. Gathering the data needed to test this hypothesis was one objective of the survey instrument.

The survey took several months to develop, including a pilot study phase during which the researcher contacted several survey respondents to get their opinions of the survey. Their suggestions for improvement were considered during the revamping process. Eventually, the instrument to be used consisted of a postcard-size survey containing five questions (see Appendix A for a copy of the survey). A letter written by the researcher requesting the subjects' participation in the survey was included with the survey. This letter made it clear that the respondents' participation in the study would be kept anonymous. The letter can be found in Appendix B. In an attempt to ensure a substantial return rate, the survey was designed to be as easy to complete as possible.

The first question on the survey asked the subjects to indicate

the total enrollment in both the largest class and the smallest class in which they taught courses during the last completed academic term. The question was:

	Largest	Smallest
How many students were enrolled in this class?	____	____

This question provided data that, when combined with responses to questions 4 and 5, allowed the researcher to note instances where class size would or would not be an issue for the professors when they decided what types of tests to use in the classes they had taught.

After indicating the enrollment for each class, the subjects were asked what percentage of their students' grades were tied to the tests in each class. This question was:

	Largest	Smallest
What percentage of the students' grades were tied to their test performance?	____	____

Since the focus of this study was on the evaluatory testing practices of college professors, this question was used to eliminate any subject who was using tests solely for diagnostic purposes, or to such a small degree that the researcher believed it to be insignificant in terms of how it would affect students' learning.

The third question asked the subjects to indicate each type of test they used in each class from a provided list:

Put a check beside the type(s) of tests used in this class.	Largest	Smallest
	____ Essay Exam	____
	____ Fill-in-the-blank	____
	____ Multiple Choice	____
	____ Oral Exam	____
	____ Short Answer	____
	____ True False	____
	____ Other (specify)	____

The list of test types in the survey does not include non-testing assessment techniques, such as term papers. Such assessment techniques were not included in this study since their effects on student learning have not been widely researched. Any attempts at assessing their effects are beyond the scope of this project.

Question 4 requested that those subjects who used more than one type of test indicate which one accounted for the largest percentage of the students' test scores. The question was:

If you checked more than one type of test in question three, which type accounts for the bulk of the students' test scores?	Largest	Smallest
	_____	_____

Responses to this question allowed the researcher to note which type(s) of tests are used most often and have the greatest effect on the students' grades. Subjects' responses to question 4 also provided information that would be used later to sort the subjects' responses into categories describing their testing behaviors.

Patterns in the testing behaviors of college professors can be illustrated by using this information in combination with data gathered from other questions. For example, were there substantial differences in tests making up the bulk of students' scores in the large and small classes?

Question 5 asked subjects to choose the type of test which, in their opinions, allows for the most accurate assessment of students' knowledge and skills. Here, they were asked to choose only one type of test. Once again, they were provided the same seven choices of test types provided in question 3 (i.e., essay, fill-in-the-blank, etc.). The question was:

> In your opinion, what type of test allows for the most accurate assessment of students' knowledge and skills?

The main purpose of this question was to provide the information needed to sort subjects on the basis of their testing practices as they relate to their beliefs about particular types of tests. Since one of the research questions asked by this study was whether a discrepancy exists between what professors choose as the most accurate assessment of students' knowledge and skills and what they use in their classroom, it was critical to know both what their preference was (question 5) and what they actually use (question 4).

Categorization Of Survey Responses

Subject responses to questions 4 and 5 allowed the researcher to divide all surveys into the five possible categories: 1)total agreement, 2)total disagreement, 3)large class agreement, 4)small class agreement, and 5)class agreement.

The "total agreement" category included those subjects whose responses indicated that the test type they selected as being the single best assessment technique (Question 5) was the same one that constituted the bulk of the students' grades in both the largest and smallest classes. For example, multiple choice test was picked as the best type of test and was also the test type that constituted the bulk of students' test grades in both the largest and smallest classes.

The "total disagreement" category was made up of the subjects who indicated that their preferred method of assessment was different from the tests they used in both the largest and smallest class they had taught during the last completed term. For example, a subject would end up in this category if he/she picked multiple choice test as the best type of test (Question 5) while at the same time indicating that he/she used short answer tests in the largest class and essay tests in the smallest class. In this case there is no agreement between practice and preference.

The "large class agreement" category was set up to accommodate subjects whose choice for the best assessment technique in question 5 matched the type of test they used in their

largest class, but was different from the type of test they used in the smallest class. A common example of the responses placed into this category was a subject who selected multiple choice as the best type of test while also indicating that he/she used multiple choice in the largest class but not the smallest class.

The "small class agreement" category described any subjects whose choice for the best assessment technique matched the type of test they used in their smallest class but not their largest class. For example, numerous subjects selected essay tests as being the best test type while also indicating that they used them in the smallest class in which they taught but not in their largest class. The large number of responses falling into this category was expected since one of the hypotheses being tested in this project was that professors consider class size as they decide what types of tests to use in the classes they teach.

The "class agreement" category was filled by subjects who indicated that they used the same type of test in their largest and smallest classes, but their choice as the best type of test was a different type of test. For example, several subjects indicated that they used multiple choice tests in both their largest and smallest classes, but selected essay as the best type of test.

These categories represent all possible combinations of responses to questions 4 and 5. In all but one instance, total agreement, some degree of inconsistency is evidenced between what college professors believe about the various types of tests, and what

they use. Subject responses to these questions helped the researcher obtain an understanding of how frequently college professors use tests in their classes they that do not believe to be best.

After all categories were developed, a sample of subjects from each category was drawn so the researcher could conduct follow-up telephone interviews of subjects from each group. These interviews were carried out for three reasons: 1)to gather information about how college professors decide what types of tests to use in the classes they teach, 2)what variables they consider when making that decision, and 3)to check the validity of the responses subjects submitted on the survey.

Telephone Interviews

After all surveys were sorted into the five categories described above, five cards were selected at random from each category. Four cards from each group were used to build the sample of subjects for the telephone interview phase of the study. The fifth card in each group was used as an alternate in the event that any of the subjects could not be located or chose not to be interviewed.

Telephone interviews were conducted by the researcher and were recorded on cassette tape. The interview dealt specifically with the decision-making process that college professors engage in when choosing the tests to use in the classes they teach. In an effort to access and make explicit the "common sense reasoning" that

subjects engage in while solving a problem (deciding what types of tests to use in the classes they teach) the researcher chose to employ a modified version of Lazzara's Interview-Analysis technique (1985).

Lazzara's Interview-Analysis Technique

Lazzara designed his interview-analysis technique as a method of acquiring expert knowledge. This knowledge normally remains "largely tacit" to the problem solver. Lazzara claims that the method is iterative in nature since prior interview analysis cycles guide subsequent interview analysis cycles. He describes the technique as follows:

> The basic GPS [General Problem Solver] methodology involves presenting the subject with a problem and asking that the problem be solved 'out loud' as he [or she] works through the problem. The session is recorded and the resultant protocol is analyzed for the heuristics that the subject used to solve the problem (Lazzara, 1985).

Lazzara developed his method of inquiry over concern with the shortcomings he believed inherent in more formal research. One of the problems with the such research is that it "predefines phenomena" and consequently, restricts the range of phenomena that are recognized, since the researcher decides, a priori, what

phenomena will offer suitable data. One consequence of prescribing data before they are collected is that this may eliminate "an essential aspect of expertise" desired, "namely, problem formation" (p. 36). Problem formation is where expertise manifests. When a problem reaches the point where it is solvable, it has emerged from a "subject-relative (or subjective) phase into an objective phase" (p.36). Lazzara claims that efforts to bring the problem "...out of subjectivity is practical, hence the importance of practical reasoning as an object of study in itself" (p. 36). Lazzara attempts to uncover the practical reasoning structures used by experts when solving problems.

In contrast to assumptions inherent in the formalist approach, "phenomenological psychology has a basic assumption that data, or better, 'phenomena,' should at first be approached without the presupposition of formal theoretical apparatus" (Lazzara, 1985, p. 36). The attempt is merely to get the subject to articulate an experience of the phenomenon in question. Phenomenologists generally claim no presuppositions about phenomena. Lazzara claims that this is impossible and introduces the idea of "multiple levels of presupposition" (p. 37). The concept of multiple levels of presuppositions suggests that limiting a subject to a particular phenomenon requires one level of distinction between what does and does not belong to the phenomenon. Lazzara claims that "the injunction that we attempt to attain an analysis that has no theoretical presumptions raises what is called a hermeneutic

paradox of interpretation" (p. 38). The hermeneutic paradox of interpretation suggests that it is not possible to understand text without a prior understanding of what it is about. A certain "preunderstanding" of the phenomenon must exist in order for communication to occur, yet "...that understanding must be altered in the act of understanding" (Winograd, 1980. pp. 223-234).

Lazzara claims that phenomenological explanations are limited to interpretive actions and thus hermeneutic in nature. In an effort to support his assertion, Lazzara cites Ihde's (1979) five hermeneutic rules to phenomenological reductions: "1) attend to what is given, 2) describe, don't explain, 3) horizontalize the phenomena, 4) seek out the structural invariants in the phenomena, and 5) correlate the poles of the intentional relation on the basis of the structural analysis" (p. 41). Lazzara contends these rules constrain inquiry.

Lazzara based his method on a study by Colaizzi (1978) which attempted to uncover the structure of experiential change that occurs when one reads a book which has a perspective changing impact. He describes the process of Colaizzi's reduction as finding themes that are thematic with respect to the other themes at that level. Colaizzi looked for themes which ran throughout the transcript. In the reduction process, "thematic structures evolve out of the constraints generated by themes at a given level of analysis" (p.49).

Lazzara also based his methodology on the works of Von Wright, who claimed that the structure of an argument consists of

"'common sense' elements, beliefs, wants, norms, and intentions" (1963). These elements provide the data necessary for the researcher to derive the steps taken when solving a problem. This method of judging behavior is said to conform to a "grammar of intentionality" (Coulter, 1979).

Harre and Secord (1972) suggest that the structure of a person's intentionality is important because it indicates "self-monitored" and "self-directed" behavior (p.9). They go on to say that self-monitoring is reflective in nature and is a mechanism which drives behavior. Such "self-direction" is based on the "...meaning ascribed to a situation" (p.6). Lazzara claims that if these mechanisms are a function of the meaning of a situation then "one very direct way of assessing the meaning of a situation for that person is by gathering first-person accounts of that person's behavior" (p. 34). Lazzara suggests that an analysis of this type of report should allow us to "...discover the relevant beliefs, wants, and intentions" of a subject's attempt at solving a problem (p. 34). He describes his methodology as one which uncovers and describes generative mechanisms.

Lazzara's procedure is made up of four stages. In the first stage, the researcher divides the transcript into segments of intended ideas which are individually listed and numbered. In any conversation, a distinction is generally made between the literal meaning of the words used and the idea intended by the speaker. This is often necessary since the speaker has several different

vehicles of communication at his or her disposal. For example, if the following statement was made, "Man, this chili is bad!," the listener would depend on the facial expressions of the speaker, along with intonation used, and consideration of where the meal is being consumed before deciding what was actually meant by the statement. In addition, speakers often formulate ideas as they speak and consequently talk in incomplete sentences, use parenthetical references, and assume a prior knowledge on the part of the listener.

In the second stage, "...each intended interpretation is placed into one of the following logical (grammatical) categories: law, norm, intention, want and belief" (p. 50). Laws include statements which note behavioral regularities, e.g., 'Everytime I come home late, my wife yells at me.' Laws also include physical regularities as well as repetitive environmental patterns.

Norms are logical or societal regularities. They exist as "... the result of convention..." rather than laws of nature (p. 63). Norms also include laws of state, laws of logic, as well as two subcategories of regulative norms: rules, and technical and hypothetical norms.

Rules are the rules which constitute a game. Von Wright (1963) declares that rules have several characteristics: 1) the rules of a game constitute the game itself, 2) rules let us know what moves are correct or permitted, 3) we understand by those moves which are prohibited as well as those moves which are obligatory.

Hypothetical and technical norms "...are more general than simple intentions..." and therefore "...considered normative for a

given individual" (p. 63). The interpretation of the generalized conditional intentions can be divided into either technical or hypothetical norms. Technical norms and hypothetical norms have the same form, but differ in the kind of condition that is mentioned as the antecedent. Lazzara, when describing Von Wright's work, declares that "the antecedent to a technical norm always mentions a "want," that is, something that is wanted, needed, or desired by the speaker" (p. 65). For example, in the following statement, "Whenever I feel exhausted, I go to the movies.," the speaker mentions a "want." In this example, the want is to rest.

The antecedent of a hypothetical norm expresses any idea, other than of a want, in a timeless framework. There is a general intention mentioned along with a general condition. For example, one could state, "Whenever it rains, I use my umbrella". In this instance, the speaker expresses an if-then relationship--when this happens, then this happens--in a "timeless" framework. No mention of a specific time is given.

The next category, "intention," is indicative of an expression of a plan of action. The speaker could mention a want but does so in the context of his or her behavior as being directed "...towards attaining the object, or state of affairs wanted" (p. 66). For example, the speaker might state, "I am going to spend the night with a friend." This statement describes intentional behavior. Lazzara states that an intention may be simple, as in the preceding sentence, or it may be more complex. A complex intention would be one

which describes a probable action, e.g., "I am going to spend the night with a friend tonight if it does not rain." In this example, the intention is probable.

The next category is "wants." The want category is closely related to the intention category. The difference is that the want expression actually states what is desired without uttering how to obtain it. For example, I could say, "I want to be able to dance like Fred Astaire." In this sentence, I express a want but no intention of how to obtain it.

Beliefs are opinions. They are the beliefs that the person hold about the world. Lazzara claims that, "the collection of 'beliefs' then, would constitute the knowledge base of the expert being represented" (p. 67). Facts are understood to be justified true beliefs.

The third stage of Lazzara's technique requires the development of a schematic representation of the steps a subject takes when solving the problem. This takes the form of a flow chart. A schematic representation is a "...representation of a person's verbal behavior, not a description of the person's 'thought processes.'"

The fourth stage consists of constructing "tree diagrams of practical arguments." This fourth step allows the researcher to search for hidden premises in arguments. Lazzara based this step on Nolt's (1984) informal argument analysis methodology. Nolt claims that the development of "tree" diagrams should follow two conventions:

1) Number the intentional interpretations that constitute a given argument.

2) Diagram the argument, using the numbers to represent corresponding statements, plus signs and lines to join premises used in a single inference, and arrows to represent intended evidential support. (p. 25)

To explicate the hidden premises in the arguments, Nolt lists four criteria to help the researcher complete missing passages in the language. The hidden premises are treated as hypotheses which are given as grounds for justification of intentional behavior. The four criteria are:

1) Add only statements which the author would have intended to be taken for granted.

2) Add premises only if they strengthen the reasoning.

3) Do not add premises which merely embody a stated inference.

4) If several different sets of premises satisfy criteria (1)-(3), add the premise or set of premises which make the argument strongest. (p. 101)

Lazzara's goal for this method of interview analysis was to provide a mechanism that would result in the development of an expert system. To accomplish this, the utilization of the labels law, norm, rule, belief, and etc. are necessary. In this study, however, the technique was used to explicate the tacit knowledge of the experts,

and to provide a template for reducing the vast quantities of data collected. No attempt was made to build an expert system. Consequently, the data reported in Chapter 4 reflect this deviation from Lazzara's technique.

Conducting the Telephone Interviews

The telephone interviews were conducted between July 1987 and August 1987. Each interview began with the researcher introducing himself and asking if the subjects remembered completing the survey. The researcher then mentioned this was the telephone follow-up to that survey and asked the subjects if they would be willing to discuss what they had noted on the survey, with their present responses being tape-recorded. In most cases, the tape was begun after the researcher had read the subjects' responses on the survey to them. This was done to refresh their memories. On a few occasions, however, the recorder was turned on during this memory-refreshing effort. Every subject approached agreed to be interviewed. Most subjects seemed very interested in talking about their responses. Consequently, no special efforts at rapport building were necessary.

To elicit a response that could be analyzed using Lazzara's technique, the telephone interviews began with the question (problem solving task),

> Think of the time when you were setting up the largest class you described on the survey. Please describe how you decided what type(s) of tests to use in the class. Please tell me what you thought about at that time.

The same question was asked later about the smallest class. Instead of asking how the subject would solve a theoretical problem, subjects were asked to recount a genuine problem (deciding what types of tests to use in both the largest and smallest classes taught). This is a deviation from Lazzara's technique, for Lazzara believes that a hypothetical situation or problem would elicit genuine verbal behavior. This researcher preferred to ground the problem in what the subject did (actual behavior) rather than to ask how they would solve a problem.

After the initial question, the follow-up questions depended on the unfolding discussion. These follow-up questions were intended to encourage the subject to adhere to the interview question and to speak specifically about their experiences in deciding on a test type rather than to discuss theoretical approaches to the decision-making process. If a subject said that she believed that multiple choice questions were flawed and the researcher noticed that she was not discussing what test type she actually used, the researcher said 'How did you decide what types of tests to use in your class?,' or 'Is that how you decided to use the test you used?' Consequently, the

interviews tended to be concise, dealing only with the original question. A sample transcribed interview is located in Appendix E.

The Personal Interviews

The final data-gathering procedure consisted of conducting personal interviews with one subject from each of the three levels of schools (small, medium, and large). The goal of these interviews was to uncover any 'deep' beliefs these professors may have had about the classroom testing phenomenon; beliefs that usually remain tacit. The interviews were also used to provide information necessary to describe how some college professors experience the testing process, and what meaning the testing process holds for them. It was hoped that combining information gathered during this phase of the study with data gathered during the first two steps (surveys and telephone interviews) would provide some insight into college classroom testing.

To gather this information, interviews had to be conducted which allowed the subjects to describe their experiences with the fewest possible constraints. Phenomenological interviews were conducted since proponents of this technique claim that they bring to the interview process the fewest assumptions and consequently circumscribe the data the least amount possible. Heidegger (1962) described the goal of phenomenological research as "let [ting] that which shows itself be seen from itself in the same way in which it

shows itself from itself," thereby allowing the researcher to describe the essence of the experience of any phenomenon (p.58). In other words, phenomenological interviews aim at letting the subject describe an experience in his or her own words without apriori constraints being placed on them by the researcher's preconceptions, including his or her methodology. The role of the researcher becomes one of trying to understand the meaning the subject is trying to convey by clarifying and exploring his or her comments. This process is believed to let the data be seen in the spirit in which it was meant. Colaizzi (1978) describes a phenomenological interview as "...a method that remains with human experience as it is experienced, one which tries to sustain contact with experience as it is given" (p. 6).

It is important to note that the data collected emerged from the interviews and were consequently limited thereby. The intent was to obtain a description of the lived experience rather than an intellectual description. May (1969) discussed this situation as the difference between knowing something, and knowing about something. To clarify his position, he described his personal fight with polio and how he experienced profound anxiety while waiting to die. He contrasted his experience with anxiety with Freud's analytical description of anxiety by claiming that "knowing" requires experience and Freud's understanding was rooted in intellect. Consequently, May labeled Freud's understanding of anxiety "...knowing about anxiety" rather than knowing anxiety (1969, p. 3). For the purpose of

this study, the researcher is interested in the lived experience rather than an intellectual description.

To conduct these interviews, the researcher randomly selected survey respondents from three schools within a 700-mile radius of The University of Tennessee at Knoxville, one from each of the three levels of schools: small, medium and large. The first three subjects asked to take part in this portion of the data gathering process agreed. The first subject selected was a professor of psychology from Berea College (Berea, Kentucky), a small school. The second interview was conducted at the University of Central Florida (Orlando, Florida), a medium-sized school. The subject was an associate professor specializing in industrial-organizational psychology. The third interview was conducted at Virginia Polytechnic Institute and State University (Blacksburg, Virginia). The subject was a professor of Educational Technology. The researcher traveled to each of these schools and spent approximately two hours interviewing the subject. All interviews were recorded and later transcribed; a complete interview appears in Appendix G.

The formal portion of the personal interviews followed an approach similar to the telephone interviews, each beginning with a single question. The personal interviews began with 'Think of a testing situation that stands out in your mind and describe it to me.' If the subject asked for more information or seemed to not understand the request, the researcher said 'Any testing experience that stands out in your mind is fine.' No attempt was made to lead

the discussion in any direction. The characteristics of what made the experience meaningful were part of the researcher's interests and therefore subjects were asked to describe what they were aware of about the experience they cited. The researcher's role was to make sure that subjects were explicit in their descriptions so the intent of their messages could be readily understood. This was done in an effort to provide the details necessary to attempt the analysis of the transcripts (protocols) later.

Personal interviews were transcribed and analyzed using a technique called "protocol thematization" (Colaizzi, 1978). Since the data were gathered via interviews, it was necessary to reduce the vast amount of information into themes which allowed the researcher to understand the meaning as intended. Consequently, a six-step protocol thematization process was employed which reduced the data into manageable units.

The thematic reduction process was adapted from Colaizzi (1978) and Kruger (1981) by MacGillivray (1985, p. 79). The six steps were:

1. Obtain a holistic grasp on the data.

2. Extract significant statements.

3. Formulate meanings.

4. Organize meanings into clusters of themes.

5. Refer clusters of themes back to original transcripts in order to validate them.

6. Formulate an exhaustive description of the fundamental structure of the phenomenon.

The intent of this part of the data gathering procedure was not to develop "an exhaustive description of the fundamental structure of the phenomenon," but rather to explore the experience of the phenomenon as a means to corroborate the other data gathered, and to formulate a tentative description of the characteristics of the phenomenon. Consequently, step 6 of the reduction process was modified.

The protocol thematization process has been used several times in recent years, with detailed descriptions of methodology appearing in MacGillivray (1985) and in Dapkus (1982). These works were based on prior works by Valle and King (1978), Romanyshyn (1977), Zaner (1964), and Merleau-Ponty (1962). Even though existential-phenomenological psychology is considered by some to be in the early stages of development, its use in these and other works suggests that it is becoming increasingly accepted as a method of inquiry.

The personal interviews were pleasant, and each of the subjects indicated an interest in the findings of the study. All subjects commented that this was the first time they had ever talked about their testing experiences, and that they found the interview to

be a learning experience for them. The subjects also noted that they were not aware of certain beliefs they had about testing until they articulated them during the interview and that the experience of doing so was a profound one.

The Pilot Study: Multiple Choice Tests

As part of the pilot study an attempt was made to determine whether the use of multiple choice tests has increased in recent years. Sixteen colleges and universities, representing small, medium, and large schools, were contacted with a request for the number of machine scoreable answer sheets that were sold or given to faculty for the sole use of classroom testing purposes. The letter was addressed to the "Testing Service Director" at the selected universities (See Appendix C). From the sixteen requests mailed, only six responses were received. The majority of schools did not have records of the numbers of answer sheets distributed in previous years. The few schools that had this information were within the large school category. There was only one response from a small school, and none from a medium size school. It was obvious that the effort required on the part of the respondents to provide this information was substantial. Since the data received was not representative of the three levels of schools and due to the effort required by the testing service directors, this avenue of inquiry was not pursued on a large scale. The responses received from the pilot

survey did, however, contain valuable information that is described in Chapter 4 (Results).

Test item bank software has recently been viewed as a convenient method for producing multiple choice tests and was seen as a medium comparable to the machine scorable answer sheets. With the prevalence of personal computers, this software is available to professors at all school sizes. An increase in the sales of such software would be another indicator of the increasing use of multiple choice tests. Thirteen software companies were sent a letter (See Appendix D) requesting a history of the number of these packages sold each year. Again, the number of responses was minimal. It has since become apparent that most test item software is being marketed by textbook publishers rather than by software companies; thus the sample targeted in the pilot study was not appropriate. Due to the poor response to these queries, they were not pursued in further detail. The data received was not considered usable for this study.

Summary

The researcher chose these methods of data collection and analysis in an effort to understand the testing phenomenon from distinctly different vantage points. Each of the three types of data gathered offered an invaluable view of differing aspects of the testing phenomenon. Sending the survey to a large number of

college professors provided a sociological perspective into current testing practices. The telephone interviews were used to extract the reasoning structures professors rely on when deciding what types of tests to use. The personal interviews were aimed at trying to understand what the experience of testing students was like for college professors.

Much discussion has recently been focused on the "politics of methodology" (Eisner, 1988). Each method of inquiry has its own inherent limitations. By using the three different methods in this research, a more complete and, possibly, a more accurate picture of testing practices in the college classroom can be achieved.

The next chapter presents the data gathered from the surveys and the interviews. Using the techniques described in this chapter, the analyzed telephone interviews are discussed with an emphasis placed on the subjects' implicit reasoning stuctures. The phenomenological interview data are presented as themes.

CHAPTER 4

RESULTS

This chapter describes the results of the study. Because of differences in the types of data, the survey results, telephone interviews, and personal interviews are presently separately. The concluding chapter draws together these three types of data to form a cohesive picture of current college professors' testing practices.

The Survey Respondents

Eleven hundred surveys were mailed to faculty at 31 colleges and universities across the United States. Five hundred and eighty surveys were returned. This represented a 52.7 percent return rate. Returned surveys were eliminated if they were too incomplete to comprehend or if there was incomplete information on the question requesting the percentage that the tests used in the classes made up the students' grades. The percentage that the test scores made up the grades of the students was believed to be a key element in the testing process since much research on the effect of tests on student behavior and learning is based upon tests having direct consequence to the students taking them. For this reason, any surveys returned which indicated that less than 10 percent of the students' grades were dependent on the tests were eliminated from the sample.

Forty-six surveys were eliminated from the total number returned, representing 0.172 percent of the total number of returned surveys and only 0.042 percent of the total number of surveys mailed. The total number of usable surveys was 534 (48.55 percent of the original mail out). The data collected by the surveys were compiled on the VAX mainframe computer at The University of Tennessee.

The largest percentage of returned surveys came from subjects representing the 'Large' schools category. This group constituted 50.73 percent of the total original mail out (1100) and 44.45 percent of the total number of usable surveys (534). The 'Medium' group made up 29.63 percent of the original mail out and 31.21 percent of the usable surveys. The remaining 19.64 percent of the original mail out and 24.11 percent of the usable surveys came from the 'Small' group. The percentage of usable surveys by each group approximates the total percentage originally mailed out to the subjects. The number of respondents by school size are presented in Table 4.

The surveys were mailed to faculty representing each academic area with a concentration in Biology, Education and Psychology; fields in which faculty should have the most flexibility in the types of tests they can use. Other faculty received surveys to maintain the ratio of school sizes. This category--'Other'--is made up of numerous academic areas including Fine Arts, Business, History, Social Work, etc.. The surveys mailed to faculty in this category

Table 4

Number of Survey Respondents by School Size

School Size	Original Mail Out		Returned		
	Freq	% of Total	Freq	% of Orig	% of Sample
Small	216	19.64	129	59.72	24.16
Medium	326	29.64	167	51.23	31.21
Large	558	50.73	238	43.27	44.45
Total	1100		534		

were selected in a random fashion with hopes of representing a variety of academic areas.

The highest number of surveys (461) were sent to Psychology faculty. Of these, 204 (44.25 percent) were returned. Education faculty were sent 422 surveys and 202 (47.87 percent) were returned. Biology faculty returned 36 surveys, 75.00 percent of the 48 sent to them. It should be noted that the 'Education' category represents several subgroups including Curriculum and Instruction, Administration and Supervision, Special Education, Educational Foundations. Table 5 describes respondents by academic area.

Table 5

Number of Survey Respondents by Academic Area

Area Taught	Original Mail Out		Returned		
	Freq	% of Total	Freq	% of Orig	% of Sample
Biology	48	4.36	36	75.00	6.73
Education	422	38.36	202	47.87	37.78
Psychology	461	41.90	204	44.25	38.13
Other	169	15.36	92	54.44	17.23
Total	1100		534		

The sample should represent an accurate ratio of all levels of professoriate across the nation. The highest percentage, 37.01, of returned surveys came from subjects who were 'Full Professors' at their respective institutions. The second highest percentage, 30.28, of the total number of returned surveys came from subjects who are

'Associate Professors' and the level of professoriate contributing the lowest percentage, 17.76, of the total percentage of the returned surveys was the 'Assistant Professor' level. It should be noted that a fourth level, 'Unknown,' was created since a relatively large percentage (14.95) of the subjects' levels were not listed in the faculty catalogs used to select them. It is possible that the subjects' levels from this category if known could change the overall ranking of the levels accounting for the total number of returned surveys. It is suspected, however, that the subjects categorized into the Unknown level represent the three ranks of the professoriate since the documents used to select the subjects made no mention of their faculty's ranks. Consequently, it is likely that the 79 unclassified subjects would represent each of the three ranks of the professoriate in a similar fashion as the 455 classified subjects. Table 6 illustrates the number of subjects who returned surveys from the three levels of the professoriate.

Table 6

Number of Survey Respondents by Teaching Level

Teaching Level	Original Mail Out		Returned		
	Freq	% of Total	Freq	% of Orig	% of Sample
Full Professor	437	39.73	198	45.31	37.01
Associate Professor	306	27.82	162	52.94	30.28
Assistant Professor	216	19.64	95	43.98	17.76
Unknown	141	12.82	79	56.74	14.95
Total	1100		534		

The Survey Responses

To provide a condensed view of the survey results, a synopsis of the subjects' responses can be found in Figure 1. To focus the respondents' thoughts on specific classes, all questions on the survey were posed first concerning the professor's largest class in the previous term, then again concerning the smallest class. Many of the respondents had only large classes and provided no data concerning the small class. It was assumed that these professors taught only one class the previous academic term. One hundred twenty-five returned surveys had information for their large class only, leaving 409 surveys with data regarding both the respondents' largest and smallest classes. The subjects' largest classes ranged from 6 students to 600. The mean for the largest class size was 74.5, the mode was 45, and the median 39. Approximately 10 percent of the respondents reported class sizes greater than 150. Twelve of these were greater than 400. These few large classes had the effect of drawing the arithmetic mean away from the median and mode and towards the few extreme class sizes. For this reason, the median of 39 seems to be the more accurate measure of central tendency of the large class size than the 74.5 mean. The small classes ranged in number of students from 2 to 452. Only five of the respondents had more than 100 students in their smallest class. There were two modes, 12 and 15; the median was 14.25, and the mean 19.

Based on the last completed academic term in which you taught courses, please provide the requested information for your class with the largest enrollment and your class with the smallest enrollment in the spaces provided below.

1) How many students were enrolled in this class?	**Largest Class** ____	**Smallest Class** ____
Range	594.00 (6-600)	451.00 (2-452)
$\overline{X}$	74.50	19.03
Mode	45	12&15
Median	39	14.25
Standard Deviation	93.65	25.86
N	534	409

2) What percentage of the students' grades were tied to their test performance?	**Largest Class** ____	**Smallest Class** ____
*$\overline{X}$	75.20	64.98
Median	79.00	65.00
Mode	100.00	100.00
N	534	409

*Only the subjects who based at least 10% of their students' grades on tests were included in this study.

3) Put a check beside the type(s) of tests used in this class.	**Largest Class** Freq	%		**Smallest Class** Freq	%
	252	47.2	Essay Exam	299	73.1
	129	24.2	Fill-in-the-blank	77	18.8
	385	72.1	Multiple Choice	169	41.3
	22	4.1	Oral Exam	43	10.5
	222	41.6	Short Answer	160	39.1
	129	24.2	True False	63	15.4
	47	8.8	Other (specify)	46	11.3
	N=534			N=409	

Figure 1. Subjects' Responses to Survey

	Largest Class			Smallest Class	
4) If you checked more than one type of test in question three, which type accounts for the bulk of the students' test scores?	Freq	%		Freq	%
	121	22.6	Essay Exam	194	47.4
	7	1.3	Fill-in-the-blank	4	0.9
	244	45.7	Multiple Choice	72	17.6
	2	0.4	Oral Exam	9	2.2
	56	10.5	Short Answer	46	11.3
	7	1.3	True False	2	0.4
	97	18.2	Other (specify)	82	20.1
	N=534			N=409	

5) In your opinion, what type of test allows for the most accurate assessment of students' knowledge and skills? (See note on back of accompanying letter.) Circle **one** option.

Essay Exam **Fill-in-the-blank** **Multiple Choice** **Oral Exam**
Short Answer **True False** **Other**(specify) _____

	Freq	%
Essay	240	45.80
Fill-in-the-blank	2	0.38
Multiple Choice	80	15.26
Oral Exam	36	6.87
Short Answer	78	14.89
True False	0	0.00
Other	88	16.79
	N=524	

Figure 1 (Continued)

Some of the large classes were actually very small (approximately 10 percent had less than 20 students), while some of the small classes were very large (5 were greater than 100). Since these types of data were not expected, no decision had been previously made to set parameters for acceptable size for large and small classes. Because two additional research methodologies were used in this study and each had their own underlying assumptions, it was determined that retaining the respondents' intended meaning was more important than the possible gains that could be made by manipulation of the data. Therefore, the data collected were not altered.

Thirty-four percent (182) of respondents use tests to determine 100 percent of their students' scores in the large class. A smaller percentage, 20.7 percent (85), base their students' entire grade on tests in the smaller class. This suggests that faculty use tests to a larger degree to determine grades in their large class than in their smaller class. Subject responses to the question about the largest class yielded a mean score of 75.2 and a mode of 100. For the smallest class the mean score was 64.98 and the mode was 100. Both measures of central tendency, in both classes, indicate a high percentage of student grades are tied to test scores. This finding allowed the researcher to base the interpretations of the data on the findings of previous research since they too had this similar characteristic.

A statistical test was used in this study to determine whether

the differences between the proportions were significant. Bruning and Kintz (1987) claim that when interested in discerning whether two independently computed proportions are different, one should use the "Test for Significance of Difference Between Two Proportions" (p. 272).

The subjects indicated that they used essay exams significantly more often in the smaller classes, 73.11 percent, than in the large ones, 47.19 percent ($z=8.1$, $p<.05$). The percentage of multiple choice tests, however, was significantly higher in the large class (72.10) than in the smallest class (41.32) ($z=9.5$, $p<.05$). These differences suggest that class size is an important consideration when professors decide what types of tests to use in their classes. Short answer tests were also used very often, ranking just under multiple choice in the small class (39.12 percent) and just after essay in the large class (41.57 percent). The remaining test types, fill-in-the-blank, true false, oral, and other (typically a combination of test types), were less popular in both the large and small classes. The rankings for the use of the test types can be found in Table 7.

Question four of the survey requested the type of test which accounted for the bulk of the students' test scores if more than one test type was used in the class. If there was no response to this question, the single test type selected in question 3 was used. The test type used most often in the large class, 45.69 percent (244) of the time, was multiple choice. In the small class, 47.68 percent (195) of the respondents used essay to determine the bulk of the

Table 7

Ranking of Test Types Used by Respondents

Large Class		Small Class	
Rank	Test Type	Rank	Test Type
1	Multiple Choice	1	Essay
2	Essay	2	Multiple Choice
3	Short Answer	3	Short Answer
4	Fill-in-the-blank	4	Fill-in-the-blank
	True False (tied)	5	True False
6	Other	6	Other
7	Oral	7	Oral

students' grades. In the large class, essay exams ranked second to multiple choice (22.6 percent). The "other" test type, usually a variety of types of tests used in combination, was used fairly frequently in both the large (17.42 percent) and small (21.76 percent) classes. Multiple choice tests were used much less often in the smaller classes to determine grades, 17.85 percent. The rankings of the test types on which the bulk of students' grades were based are in Table 8.

Table 8

Ranking of Test Types on Which Grades Are Based

Large Class		Small Class	
Rank	Test Type	Rank	Test Type
1	Multiple Choice	1	Essay
2	Essay	2	Other
3	Other	3	Multiple Choice
4	Short Answer	4	Short Answer
5	Fill-in-the-blank	5	Fill-in-the-blank
	True False (tied)	6	True False
7	Oral	7	Oral

When asked which type of test allows for the most accurate assessment of students' knowledge and skills, the largest percentage of respondents (45.80) chose essay exam significantly more often than the "other" test type which was the second choice with 16.79 percent ($z=10.13$, $p<.05$). Multiple choice was next with 15.26 percent followed by short answer with 14.89 percent. Fill-in-the-blank was selected as best by only 2 respondents, and no

subjects considered true false to be the best test type. Ten subjects indicated no preference for any test type.

The Paradox Groups

Subject responses were categorized according to the paradox schema described in Chapter 3. Since not all surveys had both large and small class data, only those with responses for both classes were used in determining the paradox information. Three hundred ninety-three surveys contained data required to determine which group the respondent fit in. Approximately 30 percent of the professors' responses at each level (full, associate, and assistant) fell into the Total Agreement category, meaning that they used the test they consider best for assessing students' knowledge and learning in both large and small classes. The percentage of professors in this category increases slightly as academic level goes up; 30.55 percent of assistant professors, 32.17 percent of associate professors and 33.58 percent of full professors were in the Total Agreement category. Since percentages for the associate and full professors were virtually the same, they were combined before being compared to the percentage of assistant professors. The results indicated that the differences between the new percentage of 32.94 (full: 33.58; associate: 32.17) and the assistant professors (30.55) percentage was not significant ($z=.2203$, $p>.05$). The highest percentage of professors in the Total Agreement category was in the Unknown group (39.13

percent). This unknown group of professors has low percentages (less than 11 percent) in each of the other categories except for Class Agreement (they used a test other than their choice of best test type in both classes), in which it has 33.33 percent. Assistant professors also had a high percentage among their ranks within the Class Agreement category, 25.00 percent. Almost the same percentage of full and associate professors' responses were in the Class Agreement category, full: 14.60 percent, associate: 13.40 percent. There was a large disparity in the number of assistant professors' responses in the Large Class Agreement category (2.78 percent) and Small Class Agreement (25.00), indicating that, except for those in the Total Agreement category, assistant professors were less likely to use the test type they consider best in their large class than in their small class.

To compare the percentages of full, associate and assistant professors in the large class agreement category, the full and associate percentages (full: 16.06; associate: 13.91) were once again combined. The percentage of assistant professors in this category (2.78 percent) was found to be significantly lower than the other two levels of professoriate ($z=2.80$, $p<.05$). The percentages of the different levels of professoriate in the Small Class Agreement category were similar: full professors: 21.17 percent; associate professors: 22.61 percent; assistant professors: 25.00 percent. One fourth of assistant professors gave a test other than their choice of best test type in both their large and small classes, putting them into

the Class Agreement category. Table 9 contains the ranking of the percentages of faculty level falling into the different categories.

Table 9

Ranking of Paradox Groups by Teaching Level

Teaching Level	Percent	Teaching Level	Percent
Full Professors		Associate Professors	
1) Total Agreement	33.58	1) Total Agreement	32.17
2) Small Class Agreement	21.17	2) Small Class Agreement	22.61
3) Large Class Agreement	16.06	3) Total Disagreement	18.26
4) Total Disagreement/	14.60	4) Large Class Agreement	13.91
Class Agreement	14.60	5) Class Agreement	13.40
Assistant Professors		Unknown	
1) Total Agreement	30.55	1) Total Agreement	39.13
2) Small Class Agreement/	25.00	2) Class Agreement	33.33
Class Agreement	25.00	3) Large Class Agreement	10.14
4) Total Disagreement	16.67	4) Small Class Agreement	8.70
5) Large Class Agreement	2.78	5) Total Disagreement	8.69

More than twice as many professors in education (45.77 percent) as in biology (22.22 percent) ($z=2.29$, $p<.05$) and psychology (21.62 percent) used the test type they consider best in both their largest and smallest classes (Total Agreement). Biology professors were more likely to be in the Small Class Agreement (33.33 percent) or Class Agreement (25.93 percent) categories. Psychology professors' responses were also most likely to be in the Small Class

Agreement category (27.03 percent) or in Total Disagreement (22.30 percent), giving two different types of test in their large and small class, neither of which was their choice of best test type. Education professors' responses were least likely (9.15 percent) to be in the Total Disagreement category. Exactly the same number of education professors' responses (16, 11.27 percent) fell in the Large Class Agreement and Small Class Agreement categories; they gave their choice of best test type in either the large or small class, but not in both. The 'Other' academic area had a very large percentage of responses in the Total Agreement category, 38.16, and the least in the Total Disagreement (13.16) and Large Class (13.16) categories. Table 10 contains a ranking of the percentages falling into the different categories.

Professors at small schools were more likely than those at medium and large schools to use the test they considered to be the best in both their largest and smallest classes. Over 41 percent of small, 30.71 of medium, and 30.86 large school faculty responses fell into the Total Agreement category. The percentage for small school professors was compared to the combined percentages of the large and medium school faculty to check for statistical significance. The 10 point difference in percentages yielded a nonsignificant finding ($z=1.95$, $p>.05$). However, given the differences evidenced elsewhere along with the inexactitude of this measure, it is asserted that the differences in this case are practically significant.

Professors at small schools were also more likely than those at

Table 10

Ranking of Paradox Groups by Academic Area

Academic Area	Percent	Academic Area	Percent
Biology		Education	
1) Small Class Agreement	33.33	1) Total Agreement	45.77
2) Class Agreement	25.93	2) Class Agreement	22.53
3) Total Agreement	22.22	3) Large Class Agreement/	11.27
4) Total Disagreement	11.11	Small Class Agreement	11.27
5) Large Class Agreement	7.40	5) Total Disagreement	9.15
Psychology		Other	
1) Small Class Agreement	27.03	1) Total Agreement	38.16
2) Total Disagreement	22.30	2) Small Class Agreement	18.42
3) Total Agreement	21.62	3) Class Agreement	17.10
4) Class Agreement	16.22	4) Total Disagreement/	13.16
5) Large Class Agreement	12.84	Large Class Agreement	13.16

large schools to use their best test choice in their largest classes (small: 17.31 percent, large: 6.17 percent). Professors at large schools used their best test choice frequently in their small class (27.16 percent) but less often in their large class (6.17 percent). Over 20 percent of large school faculty fell in the Total Disagreement category--they used a different test in their large and small class and a third test type was the one they considered best for assessing students' knowledge and learning. Small school faculty were least likely to be found in this category (8.65 percent). Table 11 contains a ranking of the percentages falling into the different categories by school size.

Table 11

Ranking of Paradox Groups by School Size

School Size	Percent	School Size	Percent
Small Schools		Medium Schools	
1) Total Agreement	41.35	1) Total Agreement	30.71
2) Class Agreement	19.23	2) Class Agreement	24.41
3) Large Class Agreement	17.31	3) Small Class Agreement	16.54
4) Small Class Agreement	13.46	4) Large Class Agreement	14.96
5) Total Disagreement	8.65	5) Total Disagreement	13.39
Large Schools			
1) Total Agreement	30.86		
2) Small Class Agreement	27.16		
3) Total Disagreement	20.37		
4) Class Agreement	15.43		
5) Large Class Agreement	6.17		

The ranking of respondents in the various paradox categories reveals that small school faculty and faculty within education have the same category rankings. The percentages within the categories are even similar, even though the total numbers for each were different (Education faculty: 142; small school faculty: 104). The highest percentage of respondents in every school size and faculty level use the test type they consider best for assessing students' knowledge and learning in both their large and small classes (Total Agreement), but the members most often found in this category are those in education and at small schools.

When considering the general picture of respondents, most responses (33.58 percent) fell in the Total Agreement category, followed by Small Class Agreement (20.11 percent), Class Agreement (19.34 percent), Total Disagreement (15.01 percent) and, finally, by Large Class Agreement (11.95 percent). In other words, approximately 40 percent of the respondents give their preferred choice in both their large and small classes. If they do not give their choice in both classes, they are then most likely to give it to their small class, or to use a different test type in both their large and small class. Fewer respondents give two different test types in their large and small classes, but consider a third to be best for assessing students' knowledge and learning. The smallest percentage of respondents gave their choice of best test in the large class, but not in the small class. The responses falling into the different paradox categories are described in Table 12.

Table 12
Paradox Categories

	Number Returned	Total Agreement			Total Disagreement			Large Class Agreement			Small Class Agreement			Class Agreement		
		Freq	%Ret	%Tot	Freq	%Ret	%Tot	Freq	%Ret	%Tot	Freq	%Ret	%Tot	Freq	%Ret	%Tot
Professor	137	46	33.58	11.70	20	14.60	5.09	22	16.06	5.60	29	21.17	7.38	20	14.60	5.09
Assoc Prof	115	37	32.17	9.41	21	18.26	5.34	16	13.91	4.07	26	22.61	6.62	15	13.40	3.82
Asst Prof	72	22	30.55	5.60	12	16.67	3.05	2	2.78	0.50	18	25.00	4.58	18	25.00	4.58
Unknown	69	27	39.13	6.87	6	8.69	1.53	7	10.14	1.78	6	8.70	1.53	23	33.33	5.85
Total	393	132		33.58	59		15.01	47		11.95	79		20.11	76		19.34
Biology	27	6	22.22	1.53	3	11.11	0.76	2	7.40	0.51	9	33.33	2.29	7	25.93	1.78
Education	142	65	45.77	16.54	13	9.15	3.31	16	11.27	4.07	16	11.27	4.07	32	22.53	8.14
Psychology	148	32	21.62	8.14	33	22.30	8.40	19	12.84	4.83	40	27.03	10.18	24	16.22	6.11
Other	76	29	38.16	7.38	10	13.16	2.54	10	13.16	2.54	14	18.42	3.56	13	17.10	3.31
Total	393	132		33.59	59		15.01	47		11.95	79		20.10	76		19.34
Small	104	43	41.35	10.94	9	8.65	2.29	18	17.31	4.58	14	13.46	3.56	20	19.23	5.09
Medium	127	39	30.71	9.92	17	13.39	4.33	19	14.96	4.83	21	16.54	5.34	31	24.41	7.89
Large	162	50	30.86	12.72	33	20.37	8.40	10	6.17	2.54	44	27.16	11.20	25	15.43	6.36
Total	393	132		33.58	59		15.02	47		11.95	79		20.10	76		19.34

Results of the Telephone Interviews

The data gathered during the telephone interviews were reduced using Lazzara's technique (see description in Chapter 3). The goal of this reduction was to identify professors' decision-making processes in choosing the tests they use in both their largest and smallest classes. To do this, interviews were scrutinized to find the subjects' base arguments for making their decision. As explained in Chapter 3, subjects generally describe only their base arguments in response to an interview question. Assuming that certain considerations are "tacit" and are not often expressed, the base argument alone is not a sufficient descriptor of the professor's thought processes. This implied information had to be identified and added back to the base argument to develop the full argument, the complete decision-making process. This was done for all the telephone interviews. The complete arguments can be found in Appendix F.

After all telephone interviews were transcribed, they were coded into argument statements. Information superfluous to the problem-solving task was omitted. From these statements, three categories were identified which provided a meaningful way to sort the responses. These categories are not an all-inclusive categorization system of the data, but they do provide a coherent system. The three categories were: 1)limitations associated with time (henceforth referred to as "time"), 2)an awareness of others

("others"), and 3)characteristics of the test and how they fit with class material ("fit").

The category of time has several characteristics. Many subjects indicated that their decision of what tests to use was made on the basis of various time constraints. For example, time as a constraint in the form of number of students in a class was mentioned by 8 subjects when talking about their largest class. They stated that a large number of students in a class made using certain preferred test types too time consuming. Phrases like "If you have 35 to 50 students in a class, it is virtually impossible to do essay exams," "My decision was made on the number of people in the class," "Mainly with that size class, something that can be graded rather quickly," "My decision was based on the size of the class," were indicative of this perception of time as a constraint. Several other experiences of time as a constraint were mentioned; time as a commodity to be spent, time as expectations by others, and time as a metric for ascertaining appropriate on-task behavior. The general tone of these descriptions was that each subject has various competing expectations of him/her and, consequently, must trade off or compromise in an effort to not spend inappropriate amounts of time on tasks. Time was mentioned 20 times for the large class and 7 times for the small class.

The second category, (an awareness of) others, was mentioned by 13 subjects when describing their large class and 7 times when describing their small class. Others can be students, colleagues or

administrators, any or all of whom can have an impact on the professor's decision to use a particular test type. Some of the statements that fit into this category were: "I use essay tests so students can qualify answers," "Students do not like taking essay exams," "Students complain about essay exams," "If you spend all your time grading papers, you will be criticized by your peers for doing it."

The final category, fit, was the second most popular category, mentioned 12 times in the large class and 6 in the small class. The subjects declared that they chose the types of test they did because their choice of test fit "the depth of the material," they were teaching at a "variety of levels" which demand a certain type of test, and they could "cover more material" if they chose a certain type of test.

Only a few statements were made which did not fit into one of these categories. The outliers were decisions based on "custom," and using multiple choice tests because the subject said it was "fun to write multiple choice questions." If the interviews had been extended, it is possible that the outliers would have been described in such a way that would have made classifying them into one of the three categories possible. Without the expanded context, the meaning beyond the obvious language could not be derived. Consequently, the two statements remained freestanding.

The following interview reduction has been included here because it epitomizes the 20 telephone interviews. The argument describes how the subject decided what types of tests to use in the

largest class taught during the last academic term. The underlined statements were added to the subject's base argument to explicate the tacit knowledge omitted. The entire interview was used along with the subject's responses on the survey instrument to build the full argument. A copy of the interview can be found in Appendix E.

> Question: I would like you to think back to when you were considering what type(s) of test(s) to use in this class and describe how you came to your decision.
>
> 1. If you have 35 to 50 students in a class, it is virtually impossible to do essay exams and that is why I always use multiple choice in the introductory class. (Rule)
>
> 2. When I can manage to survive the paperwork, I always use essay tests. (Intent)
>
> 3. One should use essay questions with a small number of students. (Rule)
>
> 4. With a large number of students, essay exams require too much time for grading. (Rule)
>
> 5. By using multiple choice questions, it is possible for me to check the validity and reliability of the test questions. (Belief)
>
> 6. Over the years you can build a test that you know is a good test. (Belief)
>
> 7. I can actually get them to do things other than memorize facts, I can get them to think. (Belief)

8. Multiple choice tests have a reputation of measuring the lowest levels of skills. (Belief)

This subject's responses to the survey placed him into the small class paradox group--his preferred test type was used in the small class, but not in the large class. As indicated in the argument, the subject chose to use multiple choice tests due to reasons other than a preference for them. His reasoning suggested that in a class with fewer restrictions (number of students) he would use the type of test he preferred. Consequently, a paradox exists between the type of test he used and what he believed to be the best test type.

This argument contains statements which fit into all three categories of responses: time, others, and fit. In a large class, the use of essay tests places too great a time constraint on the professor to be feasible. To this professor, it was important that his tests be "good" tests, reliable and measuring more than the "lowest level of skills." It was also the subject's desire to choose a test type which would "get them to do things other than memorize facts, . . . get them to think," indicating an awareness of others, the students. The subject felt that he had been able to develop a good multiple choice test over the years that fit his needs and reduced the time constraints imposed by essays, which he much preferred and would use in a smaller class. The substance of these statements was repeated in many of the other telephone interviews.

Subject comments in the telephone interviews indicate that when deciding what types of tests to use in their largest class, class size seems to be the most salient variable considered. If class size seems manageable in terms of time constraints, professors tend to use the types of test they believe to be most appropriate. Appropriateness is defined as how well the test fits the class material in terms of depth and amount to be covered, and the effects the test has on others--students, other faculty, and administrators.

For the smallest class, the number of students in the class does not appear to be as important in the decision-making process as it is in the large class, except in those cases where professors teach only large sections. The subjects indicated on the surveys and in the telephone interviews that they tend to use tests they believe to be best in the smallest class to a much greater degree than in their largest class. Other variables considered along with size of class are the level of the class, i.e., freshman, senior, or graduate class, and the objectives for those courses. These variables seem to be more predictive of test types used than class size alone.

Subject responses to the surveys helped to categorize them into the paradox groups described in Chapter 3. Tables 13 through 17 contain the pertinent survey data and the primary ideas of the subjects' interviews, sorted by paradox group.

Table 13

Telephone Interviews in Total Agreement Paradox Group

ID	Size	Area	Level	Type Used Lg	Type Used Sm	Best Test	Comments Large Class	Comments Small Class
060M13P2	M	Psy	Assoc	E	E	E	Essay exams require the students to learn the information, not just memorize it.	Essay exams show me how well they learn the material.
266M20E3	M	Ed	Assis	E	E	E	Essay exams allow students to express themselves and develop their writing skills.	I have a philosophical orientation toward developing and continuing an ongoing dialogue with these graduate students and essay exams allow me to do that.
206M18E2	M	Ed	Assoc	MC	MC	MC	Multiple choice tests should be used since students don't like essay questions and since information can be covered as well with multiple choice.	Students do not like taking essay exams and the information can be covered as well with objective exams.
202S30O5	S	O	U	MC	MC	MC	Multiple choice tests provide a common language between teachersand students and provide ample feedback to students.	Multiple choice exams can be graded quickly and can provide feedback to the students.

Table 14

Telephone Interviews in Total Disagreement Paradox Group

ID	Size	Area	Level	Type Used Lg	Type Used Sm	Best Test	Comments Large Class	Comments Small Class
481L09O2	L	O	Assoc	MC	O	E	Objective exams allow the instructor to get them graded quickly and back to the students, thus allowing more time for valued activities.	Essay tests allow students to improve their writing skills.
306M21P5	M	Psy	Unkn	MC	O	E	Decision is based on size of the class as well as a desire to cover a lot of information.	I wanted to probe for depth of knowledge and essay and oral tests were best for that.
181L03P1	L	Psy	Full	O	SA	E	Even though essay tests can benefit the students in a variety of ways,multiple choice must be used because of the time required to grade essays.	The type of information tested determines which type of test will be used.
307M21P5	M	Psy	Unkn	MC	E	O	Objective tests are used in order to cover a large amount of material in a reasonable amount of time.	Objective tests are used in order to cover a large amount of material in a reasonable amount of time.

Table 15
Telephone Interviews in Large Class Agreement Paradox Group

ID	Size	Area	Level	Type Used Lg	Type Used Sm	Best Test	Comments Large Class	Comments Small Class
135S30E5	S	Ed	Unkn	MC	E	MC	Multiple choice exams are used because of ease of grading.	Essay exams allow for better evaluation of the students' learning.
206S30O5	S	O	Unkn	E	MC	E	I select my tests on the basis of class size, level, and the kinds of things the students are doing.	I select my tests on the basis of class size, level, and the kinds of things the students are doing.
012S22B1	S	Bio	Full	SA	E	SA	Students benefit from more than one type of test question than another and less time is involved in grading tests that are not strictly essay.	Supply tests teach the students to think critically.
080M14P2	M	Psy	Assoc	MC	E	MC	The test allows better evaluation of the students' learning.	My decision was based on time limitations.

Table 16
Telephone Interviews in Small Class Agreement Paradox Group

ID	Size	Area	Level	Type Used Lg	Type Used Sm	Best Test	Comments Large Class	Comments Small Class
011L01P3	L	Psy	Assis	MC	E	E	Since essay tests are too time consuming to grade for a large class, the size of the class determines the type of tests used.	Essay tests require the students to demonstrate that they have worked with the material.
190M17E2	M	Ed	Assoc	MC	E	E	Multiple choice exams are the most appropriate way to test certain types of information.	The multiple choice exams are the most appropriate way to test certain types of information.
134S30P5	S	Psy	Unkn	MC	E	E	The objective test is the practical thing to use in a large class and, over the years, one can build a good objective test.	Essay tests are used to make the students think and relate things to each other.
002L01P1	L	Psy	Full	MC	E	E	Multiple choice tests are most easily graded.	The testing process should be as time efficient as possible so as not to take time away from other duties.

Table 17
Telephone Interviews in Classes Agreement Paradox Group

ID	Size	Area	Level	Type Used Lg	Type Used Sm	Best Test	Comments Large Class	Comments Small Class
305M21P5	M	Psy	Unkn	MC	MC	SA	I have confidence in my multiple choice tests' ability to measure what I desire and they are the most convenient test type.	Multiple choice exams are used for expediency.
484L09O2	L	O	Assoc	MC	MC	SA	Class size dictates the type of test to use and, although not the best test type, multiple choice can be effective and give the best and most timely feedback to students.	While class size is a limiting factor, it is necessary to use supply tests to be sure students are learning the right things.
215S31E2	S	Ed	Assoc	SA	SA	MC	Essay exams fit the needs of this class better than objective exams.	Objective exams would not help to achieve the goals of this course.
065S26E2	S	Ed	Assoc	SA	SA	E	When the class is large, objective tests should be used, conversely, when the class is small, short answer tests should be used since they allow the instructor to see attitude changes toward the material.	I can better evaluate student learning with supply tests than recognition tests.

Results of the Personal Interviews

The data gathered during the three personal interviews were reduced using the six step thematic reduction process described in Chapter 3. Superfluous text was carefully removed from the transcribed interviews (protocols), once as the researcher read them to "obtain a holistic grasp on the data," and a second time in order to select statements significant to the subject's experience of the testing process (the "Second Reduction," see Appendix H). Completion of this step resulted in a list of each subject's most significant statements, rephrased to state only the idea their full statements conveyed, without unnecessary words or redundancies. The rephrasing often incorporated the phrase 'I am aware of...' into the subjects' descriptions. These statements were analyzed to find underlying themes, resulting in the "Thematic Reduction." For the final steps of the protocol analyses, themes with similar meanings were grouped together and validated against the original text of the protocols. The final result provides a glimpse of "the structure of the phenomenon" of how college professors experience the testing process. Information about subjects, interview duration, and number of significant statements can be found in Table 18. Representative significant statements culled from the original interviews are listed in Figure 2.

The themes that emerged during these reductions were similar to the categories developed from telephone interviews: time, others,

Table 18
Significant Statements in the Personal Interviews

Subject	School Size	Duration of Interview	Number of Significant Statements	Paradox Group
1	Small	85	22	Small Agreement
2	Medium	83	26	Small Agreement
3	Large	88	20	Total Agreement

1. Time needed to go over the test.
2. Time constraints associated with a large class.
3. The importance of grades.
4. Inherent subjectivity in grading.
5. Students' opinion of test.
6. Testing not being the end of the teaching process.
7. Class size.
8. Topic.
9. Importance of constructing good tests.
10. Time needed to take the test.
11. Students need to express themselves on exams.
12. Need to base exams on stated objectives.
13. The level of students.
14. Tests as a teaching device.
15. Tests as a measure of teaching effectiveness.
16. Tests as a method to encourage good study habits.
17. The need for quick feedback.
18. The possibility of students cheating.
19. The testing environment.
20. Time required for grading essays.

Figure 2. Typical Answers to the Question:
What Are You Aware of When Testing Students?

and fit. A fourth theme was found in the personal interviews that had not been mentioned in any of the telephone interviews--student grades. From the personal interviews it was obvious that the theme of grades is an integral part of the experience of testing.

The themes which emerged during the early stages of the reduction process appeared against the backdrop, or as Gestalt psychologists describe it, the "ground" of the teaching/learning process. Subjects described the process as cyclical, consisting of the following stages: design the course, teaching the material, administer the tests, grade tests, go over tests, teach new material, and so forth. It is against that ground that the subjects' descriptions of their experiences of testing make sense, or emerges and becomes "figural" for them. Attempts to decontextualize the figure from the ground render the experiences meaningingless.

While the testing process is generally held to encompass only the administering and grading the test, subjects in this study describe their experience of testing as being inextricably imbedded in the overall process of teaching/learning. When asked to describe their experience of testing, each subject spoke of building the test, administering it, grading it, going over it, and then going on to new material. As they discussed their experience of testing, four themes emerged--others, time, fit, and grades.

The Awareness of Others in Testing

The first recurrent theme, cited by all three subjects, concerns the professors' awareness of others. Primarily these subjects are aware of their students, but colleagues were also mentioned as they discussed their experience of testing. Professorial awareness of students was expressed in four distinct areas--student opinions, assisting the students in learning, the testing environment, and cheating. The rephrased statements included are examples taken from the second reduction phase of the interview analysis (see Appendix H). The first number identifies the subject who made the statement (1-Small School Subject, 2-Medium, 3-Large) while the second number represents the statement number in the second reductions.

The professors experience students' opinions as part of the testing process. One professor notes that his students like the tests he gives.

> 2-13 I am aware of the students' opinions of my tests. They liked taking two parts of the test in one night. It's a big occasion, but I think they liked it.

> 2-15 I was aware that the students who did well in both areas seemed to like the test.

Although it may not affect testing practices, students often

complain about tests. Two of the subjects deal with student complaints by teaching them about testing. Fostering an understanding of the testing process among the students helps to alleviate unjustified complaints.

> 2-19 I like teaching about testing in part because I am aware of how much undergraduates like to complain about tests. I tell them, "you're going to know about reliability and validity and when you want to complain about a test now, you can use some words that'll intimidate the faculty members--value analysis, p-value, etc."

> 1-18 When we go over the test, I am aware that the students need to know that I am really interested in whether they understand the material and that I see going over the test as another learning experience. If there is controversy over an item, we discuss how to decide if an item is ambiguous or just hard. They like that and they go and bug the other teachers about it.

One reason students may complain is that they do not believe they are treated fairly by the testing practices of other professors. As the two subjects note above, the students can take their new-found knowledge of testing and "bug the other teachers about it." Fairness to the students is an important factor in the testing process.

3-5 When I give a test, I am aware of the way it is based on the objectives I have given the students. The students weren't shocked; they knew what was coming in the test. I just asked for details.

3-13 I am very aware that tests should be built around stated objectives. You know what you want the students to do as a result of your teaching. I give detailed instructional objectives beforehand, then try very hard to build my tests around them. I don't like the concept of testing as punishment which it would be if there were unexpected questions.

2-20 I am aware of students' opinions of our tests. In our evaluations, students rate the item that says "test evaluations are made in a fair and impartial manner" below average at all levels, individually, department, college, university. You can just see the whole pattern, the profile.

The subjects are also aware of the students' preference for certain test types. They may or may not, however, elect to accommodate the students' preference.

2-14 I am aware of the students' opinions. In my undergrad statistics class, they don't want concepts tests because we cover more topics in a wider, less detailed manner.

2-18 I am aware of the subjectivity of essay tests. Students have said to me that they do better

> on essay tests and I think "because you get away with bullshit too often, and I'm not going to let you do that." I think overall they probably like objective tests better--they are easier to take, you don't have to generate or start your own response.

The need to use different test types is sometimes predetermined by the level of students in a class.

> 3-6 In a testing situation, I have to be aware of the level of student. This was a 6000 level graduate class, a very self-selected group. I've never felt comfortable giving an objective test at that level.

One subject notes an awareness of a difference in students' reactions to the test. Controversy over multiple choice items elicit group challenges, but essay items are questioned individually.

> 1-9 I am aware of a difference in the students' attitude toward the multiple choice items and the essay. In the group, students will ask questions and challenge the multiple choice more. But the essay becomes a private thing; students come up after class and say, "I don't understand."

Professors also experience the testing process as a method of assisting students in learning. The use of essays allows for unique interpretations of the same question, as opposed to the limiting

provisions of a multiple choice item. The freedom to express themselves fully must be tempered with a focused response.

> 3-4 I am aware of the students' need to express themselves fully on an exam. Objective tests do not allow a student to express himself or get out as much information as he wants to. I feel very strongly that people ought to be creative about this. They may know some information my question doesn't quite get at, and being able to explain their rationale seems a healthy thing to me.
>
> 3-10 Even though I encourage students to tell me why they're thinking a certain way, I am aware that their answers can get out of hand. The answers get a little bit longer than you like. I've tried all kinds of ways to force them to be more concise, but I don't want to defeat my other purpose. . .to expand.
>
> 1-6 With incoming freshmen, I discovered that they haven't learned to focus on the bigger issues, so I give them a set of essay questions beforehand, one of which will be on the test. My idea is to use the essay questions to give them a little focus, and then use their discipline to figure out the essay questions in advance. I am very aware of their need for direction and how I can use the test to teach them discipline.

One professor views the testing process as a means of helping the students build good study habits, particularly lower division students.

1-4 In the lower division courses, I test the students frequently. I am aware that the preparation for the tests helps to build good study habits.

1-5 I am aware of my responsibility to give the students a chance to learn how to study. Many of them have done well enough in high school without having to learn any discipline. Our school is a busy, demanding place and my idea is to instill a consistent discipline with frequent testing.

1-19 I am aware of testing as a method of building skills. I view it as a structure to enhance the whole process. After the first test, I talk to the students about how they did and what they can do to improve.

Along with good study habits, the test should be used to give students the ability to think abstractly, something they are not always willing or able to do.

1-3 I am aware of the difficulty of making students engage in some sort of higher level mental operation; to look at the implications of factual data and think about it. In one class we had discussed the sensory projection areas of the brain and their size in relationship to the sensitivity of their respective body parts. The multiple choice item said "Which part of an elephant's sensory projection area would be larger, that which came from the tip of his trunk, or that which came from the back of his leg?" One of

> the students was very irritated and said, "You never said anything about elephants in class." What struck me was the total unwillingness to deal in the abstract issues, to draw from the sheer fact.

The subjects experience the testing environment as having an important effect on their students. All subjects mentioned the need for the students to have a relaxed testing environment.

> 2-17 I am aware of the demands of test taking, so I always tried to create a nice testing environment where they wouldn't feel like I was trying to harm them in some way. They'd laugh about something even while they were taking the test. The environment was good.

> 3-11 I am aware of the testing environment. Our school has an honor code system. We don't have to be there or monitor a test. The students know where I am if they have a question during the test.

> 1-13 On test day, I am aware of the tension the students feel. I try to reduce that by telling a joke. They have come to expect it.

The experience of the testing environment was very different for the two subjects who discussed cheating. One subject felt no responsibility to monitor his students during test taking because of his confidence in the school's strict honor code system. The other

subject, however, expressed distress over cheating and was unhappy in having to deal with cheating as part of the testing process.

3-12 I am comfortable with giving tests at the students' leisure because of the strength of the honor code. I give them a sheet to sign that they pledge that they have neither given nor received help on the exam. And with the honor code, if they sign the sheet, that's alright. There are strict penalties, so I think it works. There will always be people who cheat no matter what.

1-14 I am very aware of the students' nonverbal behavior during test taking. I walk around the room, just to make my presence felt and I tell them to raise their hand if they have a question and I'll come to them, rather than disturbing the others.

1-15 Part of my awareness of their nonverbal behavior is focused on the possibility that they are cheating. It distresses me and is hard to deal with when I see a student trying to look at another student's paper. I realize that a lot of times during a test you kind of wander around, so I keep an eye on the student to see if it was just a vague behavior or actually cheating. I feel strongly about it.

1-16 I had to confront one student about cheating and it was upsetting. I can't understand how anybody would do that, and you know they're cheating themselves more than the system. It takes me a while to verbalize my energy to stop them. I think that, since they

> are kids, we go through everything we can to help them succeed.

An awareness of colleagues was mentioned infrequently by the subjects. When discussing students complaints about tests, subjects 1 and 3 noted that the students can use the information about testing to "bug" or "intimidate" other faculty members. Subject 3 is also aware of his colleagues' criticism of his college.

> 3-19 I think it is my responsibility to hold the students accountable for what I said they would be. At the same time, colleagues say they give easy grades in Education. My retort is "Come in and sit with my class. See how easy it is." I am aware that my primary responsibility is to the students.

Others, primarily students, are experienced as an important part of the testing process by all subjects. They are aware of the students' opinions and a desire to help the students learn, from allowing self-expression through essays to providing a relaxed testing environment.

The Experience of Time in the Testing Process

Subjects experience time as a constraint, affecting them either within the confines of their class or outside it. They acknowledge the effects of time in a variety of ways.

2-3 I am aware of the time it takes to build a test. It depends on whether you have to start from scratch. You can cut and paste and revise from older tests, and that isn't nearly as painful as starting from scratch.

3-3 Sometimes I wonder why I give essay exams when I am aware of the time required to grade them. I tell the students to expand and to write what they want and make sure they come to the point. Later I am sitting there with three days to get grades in and 20 stacks of essay exams. Then I wonder why I did this.

1-2 I feel strongly about giving the test back the next day. The first test has 20 multiple choice items and two essay questions. I have a machine score the multiple choice and I hand-score the essays. I am aware of the need for quick feedback.

The subjects mention the effects of time as a constraint outside the confines of the class less often than within, but there was an awareness of it.

2-2 I am aware of the time required for the type of test I plan to give when there are so many other things I need to get accomplished. I've thought of giving a shorter test, but I rarely do.

Each subject appears to have a preconceived notion of an appropriate amount of time that should be spent engaged in the

different facets of the teaching/learning process. As they experience the steps in this cycle, professors become aware of time as a commodity. Their notions of appropriate time frames become figural for them. It is reasonable to assume that the subjects have the same perceptions of time which should be spent on their job duties outside class. When either inside or outside of class time requirements exceed what they expect, compromises must be made to accommodate the limited time available.

> 2-4 The size of the class makes me aware of how I can use the test. If I had 55 or 60 people in there, I might give them the same kind of test, but I might not spend as much time going over it with the idea of teaching from it. What is the practicality in beating yourself to death by trying to be fair to the students?

> 1-1 I am aware of class size when I decide what test type to use. If I had sections of a hundred, I wouldn't attempt essay questions. With thirty, I find I can do the multiple choice and essay, but it does make it rough.

Fit with Course Material and Objectives

The theme of 'fit' is indicative of several different experiences, one of which is the relationship between the test and course material. The subjects indicated that tests should mirror the

material covered in class, that certain test types are more appropriate for some classes and for evaluating learning. They experience testing as a means of accomplishing certain goals, provided there is the necessary 'fit' with the course material.

> 2-25 I am very aware of the need for the content of the test to reflect the things you've been doing in instruction--topicwise. I may take a limited amount of material out of the reading assignment, but the primary content of the test will come out of things we actually discuss in class.
>
> 1-20 I am aware that certain test types are more appropriate for some courses. In a survey course, I use multiple choice because I want to make sure they're getting something from their reading. Other courses, I use short answer and essay.
>
> 1-22 When I am creating a test, I am aware of what I want each item to do as far as evaluation goes. I use multiple choice to see if they are reading and comprehending the text. I try to vary the difficulty and the type of questions the tests have. I keep records on the mean scores on the multiple choice and the difficulty level to get some idea about the reliability.

The level of students in a class also affects the type of tests that the subjects use. The subjects indicate an awareness of graduate and undergraduate students operating at differing levels

which require different testing strategies to bring about appropriate learning.

> 3-6 In a testing situation, I have to be aware of the level of student. This was a 6000 level graduate class, a very self-selected group. I've never felt comfortable giving an objective test at that level.

> 3-7 Because of my awareness of the student level in my courses, I don't believe that undergraduate students have a lot of background and information that they can bring into a test. I use short essay to give them the opportunity to expand.

The goals subjects have for a course are not tied to the course material. A more general teaching agenda encases specific goals for matching test to course material. Two subjects experience the use of tests as an effective teaching device. The review of the test provides an important opportunity for additional instruction.

> 1-21 I enjoy going over the test, because I can use the items to talk a little bit more about the material. I am aware that the test does not have to be the end of the learning process.

> 3-14 I don't object to the tests. I am aware that I can use them as a teaching device. I like to come in and say, "Okay, I'll have a test today," and then go over which points they got and didn't get.

2-22 I am aware that the test is not the end of the teaching process. It's easy to give tests back, but I think they need to be restated, rehashed. The students don't always understand that I want them to learn the material. They will say, "why don't you just give me my grade?"

Tests are not experienced solely as a benefit to the students. Professors are also aware of a test's ability to evaluate the level of effectiveness of their instruction.

2-21 I am aware of testing as a teacher effectiveness tool. Sometimes when you look at a test you've given, you get some clues to your instructional deficiencies. When certain items are missed by a sizable number of people, or even just a few, it's worthwhile to go over that topic again. One more shot at instruction. I think it's a real good use of tests--after evaluation, continue instruction based on the responses to the test.

3-15 I go through a topic and I think I've done well and when the students get to the lab, the can't even get the thing turned on. So sometimes you have to use other ways to see if you're getting the point across. I like to use tests for that. If you give the students your objectives and you teach and they don't get them, you can't say "That's a bunch of dumb people." You have to say, "What did I do as a teacher that was wrong?"

The testing of college students is not experienced as the culmination of the teaching/learning process. Rather, it is a useful method of continuing the students' instruction and of evaluating the professors own effectiveness. The test must be able to provide a vehicle for appropriately testing course material, continuing the students' education in the material, and evaluating the professors' instruction.

The Experience of Grades in the Testing Process

During the testing process, subjects are aware of the relationship between their test and grades. While subjects collectively experience testing as a positive teaching tool, grades are not necessarily accepted in such a favorable light. One subject comments:

> 3-17 I love teaching, but the thing I have never liked about it from the first day is grades. I'd really be in seventh heaven if I could teach and interact and do my thing and then when grade time came, not have to do it. I am aware that they're necessary, but I don't give them lightly and I spend a lot of emotional energy on grading.

Although grades may not be regarded positively, the subjects are aware of the significance attributed to grades, and do not

attempt to disparage them in any way. Grades are experienced as an important part of the teaching/learning process.

> 2-9 I am aware of the significance of students' grades. I wasn't about to use a quantitative class to drum people out. On the other hand, there's a notion that it's sometimes pretty easy to get A's in graduate courses. I want it to be a little bit challenging.

> 3-18 I don't like giving grades because I am aware that no matter how quantitative you try to be, there is always a subjective judgment. People consider grades important and I don't want to make mistakes. I don't want to deny a guy a grade because of three points.

Grades do not only effect students. Professors are very aware of the opinions of others in respect to their grading practices.

> 3-16 I am aware of my reputation as an educator. I do not like the idea that a lot of good grades mean you're easy or not a good teacher, or that a lot of bad grades mean you're a good, tough teacher. The College of Education has gotten a bum rap on grade inflation, but I think it's because we take testing more seriously. I let the students know what I am going to hold them responsible for. As a result, they generally do well.

The subjects experience grades as a way of rewarding their more capable students. They are aware of the grades' effects on students and the need to be accurate.

> 2-8 I am aware of the differences in students' grasp of the material. The ones who could work the problems and master the concepts I thought ought to be rewarded with a really good grade.

> 2-10 I am aware that doing well on a tough test can be rewarding, but I don't want to punish people who didn't quite handle it all.

Grades provide not only a way of rewarding the good students, they should also reflect the caliber of student. The subject quoted below implies that a recognition of the differing levels of ability in graduate and undergraduate students result in different standards for grading.

> 3-20 When I'm working with a graduate class, I am aware of the quality of student at that level. They are self-selected, a dedicated group of people. I don't think at the graduate level we work on a normal curve. I think it's a very skewed curve. I don't believe in grading on a curve. If a class of 10 people all did terribly on a test, there are going to be 10 F's. If they deserve an A, that's what they're going to get.

Perhaps one reason for the subjects' dislike of giving grades is their awareness of an inherent subjectivity in the grading process. The subjects imply that, regardless of their efforts to use objective tests, grades cannot be viewed as being very objective.

> 2-18 I am aware of the subjectivity of essay tests. Students have said to me that they do better on essay tests and I think "because you get away with bullshit too often, and I'm not going to let you do that." I think overall they probably like objective tests better--they are easier to take, you don't have to generate or start your own response.

> 3-18 I don't like giving grades because I am aware that no matter how quantitative you try to be, there is always a subjective judgment. People consider grades important and I don't want to make mistakes. I don't want to deny a guy a grade because of three points.

> 1-8 I am aware of the subjectivity in my essay grading, but I am comfortable with it and the students accept it. You get a nine if you've got everything there and a ten if you got it there in good form.

Professorial awareness of grades is an important aspect to their experience of the teaching/learning process. While grades are viewed as an important part of the process, their perceived subjectivity is a matter of concern to all of them.

The personal interviews provided a valuable insight into the

subjects' experience of the testing process. The themes expressed were similar to those found in telephone interviews, although the subject of grades was not mentioned in response to the request "How do you determine what type of test to use in your classes?" Grades appear to be a notable part of the experience of testing, but not necessarily affecting the professor's decision to use a certain test type. Time as a constraint, an awareness of others, and fit with material and course objectives were three themes occuring in the subjects' experience of the testing process described in the personal interviews and in the telephone interviewees' descriptions of how they choose a test type.

Use of Machine-Scorable Answer Sheets

A history of the number of machine-scorable answer sheets (MSAS) distributed to faculty for the sole purpose of evaluating students was requested from sixteen small, medium, and large schools as part of the pilot study. From the sixteen requests, only six responses were received. Because of the poor response and the extreme effort that was required on the part of those who did respond, a large scale survey of schools was not conducted. The information acquired through the pilot study, however, was helpful.

Six schools, five large and one small, responded with a history of the distribution of MSASs. The one small school respondent indicated that a history of the number of answer sheets was not

Table 19

Number of Machine-Scorable Answer Sheets at Five Large Schools

Year	School A	School B	School C	School D	School E
1967	89,620	225,401	•	•	•
1968	103,090	209,632	•	•	•
1969	96,874	174,173	•	•	•
1970	66,637	157,329	•	•	•
1971	111,302	136,132	•	•	•
1972	84,354	133,151	•	•	•
1973	142,312	152,755	135,051	•	•
1974	119,639	171,089	174,052	•	•
1975	152,239	195,187	200,137	•	•
1976	46,073	209,188	201,356	•	•
1977	148,742	211,018	205,893	•	•
1978	105,900	189,588	293,532	•	•
1979	104,333	171,434	350,896	•	•
1980	80,604	201,470	466,667	•	•
1981	115,083	205,355	501,573	387,000	499,024
1982	118,476	175,722	520,631	402,000	461,577
1983	155,566	170,558	547,365	451,000	468,247
1984	118,861	155,936	590,393	426,000	545,563
1985	134,430	169,253	697,836	475,000	552,134

available, but that the number had increased from 10,000 in 1985 to 15,000 in 1986. The large school respondents reported 20, 13, and 5 year histories, depending on the availability of their information (see Table 19). The histories indicate a general increase--dramatic in school C--but some of the schools show no particular trend in MSAS usage, with numbers increasing, decreasing, and increasing again from year to year. As a possible explanation for this, one test director (from school A) indicated that, although his school had kept detailed records, the pattern of MSAS usage was disguised by shifts in record keeping procedures and by a shift from a centralized accounting system to a deparmentalized one. Consequently, he described using MSASs as a very common practice, to a much greater degree than reflected in the numbers. The data presented here should be considered within that context.

The final chapter will use the data presented in this chapter to answer the questions posed by the study. The implications of the findings and the limitations of the study will also be discussed.

CHAPTER 5

DISCUSSION

From the beginning of this project, one objective was to understand what is going on in college classrooms across the country regarding the testing of students. To accomplish this, it was necessary to understand several facets of testing. Consequently, six research questions were developed which seemed to hold the key to understanding. To answer the questions three different procedures were used: a survey, telephone interviews, and personal interviews. From the data gathered on the basis of these techniques, the following can be surmised.

College professors across the United States use a wide variety of different types of tests in the courses they teach. In their largest classes, which average approximately 45 students, they use multiple choice tests most often. In their smallest class (average of 16 students) they use essay tests most often. Slightly more than two-thirds of all college professors do not use the test type they believe to be the best in one or more of the classes they teach. Differences exist between the testing practices of professors based on the size of school which employs them, the academic area they represent, and their rank. Decisions to use certain types of tests are based on the professors' perception of time as a constraint, an awareness of others, and the way the test fits with class material.

Professors experience testing against the backdrop of a more complexly patterned teaching/learning process. This process provides a ground for their experience of the themes identified as part of their decision-making process (time, others, fit) as well as the relationship of testing with grades.

Test Types Used in the College Classroom

The first question under consideration in this study was, 'What types of tests are used by college professors?' Survey data clearly indicated that a variety of test types are being used in college classrooms. All the test types listed on the survey were chosen as being used in both the largest and smallest class. In the largest class, the most popular test used was multiple choice, followed by essay, then short answer. Multiple choice and essay were also the test type which account for the bulk of student grades. There were sizable differences in the percentages of respondents who use some of the test types compared to the percentages of tests which which they reported as making up the bulk of students' test scores. For example, 41.6 percent of respondents indicated using short answer tests but only 10.5 percent claim they make up the bulk of students' scores. It should be noted that part of the differences can be attributed to the fact that the subjects were at liberty to select several types of tests on question 3 (types of tests used) and only one choice on question 4 (test which makes up the bulk of students'

scores). These differences, however, suggest that a test type hierarchy exists--professors use a variety of test types but exploit certain favorites. This was discussed by several subjects during telephone interviews. This phenomenon is made even more clear when one considers the survey question concerning the test type the subjects believe to be the best. In short, it can be said that college professors tend to rely on multiple choice tests in their largest classes. This seems reasonable since the average (mode) enrollment for these classes is 45, and since most faculty teach more than one course. The implication that pragmatic concerns affect the decision of which types of tests to use emerged numerous times in the data and will be discussed later.

In the smallest class, essay, multiple choice and short answer tests were the most often used types of tests in that order. As in the largest class, large differences exist between the percentages of the tests used and those which make up the bulk of the students' scores. Professors use a variety of test types in their smallest class with essays easily the most popular choice.

Is Multiple Choice Test Usage on the Increase?

To answer the question 'Is the use of multiple choice tests on the increase?' it was originally planned to gather data from two sources, college testing services and software companies. In a pilot study, letters were sent out requesting information from numerous

companies specializing in computer test item bank software and from the directors of sixteen college and university test centers. These avenues of exploration were dropped from the study because of the poor response rate and the obvious degree of effort required by respondents to complete the task. Consequently, the data in Chapter 4 are incomplete and unable to provide an answer to this question. There are, however indicators which do shed some light on the question. For example, almost all schools returning information indicated that many thousand machine scoreable answer sheets (MSAS) are used every year by their faculty, and that the numbers have risen, in some cases substantially. At one university, 697,836 MSAS were in used in one year. This number probably exceeds the total number of MSAS used in the entire nation twenty-five years ago.

Although responses from the software companies provided virtually no data, informal evidence in the form of conversations with numerous book publisher representatives suggests that it has become a "huge" selling point for textbooks to come with their own objective test manual and computer test bank software. Several representatives claimed this packaging is a relatively new phenomenon which has revolutionized their business. One representative claimed that she does not bother trying to sell psychology textbooks without accompanying computer software. This proliferation of personal item banking software is similar to the explosion of MSAS when the original scoring machines came into

existence in the late 1950's and early 1960's. This decentralization of item banking software from testing centers to departmental and even individual professors offices is one explanation for the lack of uniformity in the five schools who reported on their use of MSAS. It was noted by one test director that their record keeping procedures had in fact shifted from a centralized location to a more departmentalized system.

The following is an example of the far reaching implications of the competition between book publishers which has had an impact in the college classroom testing pratices of professors. A general psychology textbook was recently adopted by a university with an enrollment of 10,250 students. One of the main reasons for the book's adoption was that it came with a large multiple choice test builder; one with several thousand questions. Each quarter at this university approximately seven sections of general psychology are offered with an average of 150 students in each. What the adoption of this book means is that around 3150 students will be tested yearly using multiple choice tests in this one class. With the accepted use of these test builders, one can begin to appreciate the degree to which multiple choice tests are growing in popularity.

The substantial numbers of MSAS indicated by the respondents, the pervasiveness of test bank software at the individual professor level, and the standardized testing emphasis in our public schools suggest that objective tests, more specifically, multiple choice tests have become the test of choice and are being

used to a greater degree today than in previous years. The degree to which the survey respondents use them in their courses also adds support to this claim. Almost three fourths of the survey respondents used multiple choice tests in their largest class and more than 40 percent used them in their smallest class. Although limited in its generalizability, additional support comes from the telephone interview data. Several of the subjects suggested that their decision about which tests to use was influenced by, if not based on, pragmatic concerns related to job expectations. They further implied that they are feeling more and more pressure to engage in a variety of activities as part of their obligation to their institutions. These competing activities impinge on the time they believe they can afford to expend in the testing process. Consequently, they feel the need to test in the most efficient way, which often translates into using multiple choice tests.

Do Professors Use the Test Type They Consider Best?

The third research question explored by this study was, 'Do college professors use the type(s) of tests they believe to be the best method(s) for assessing students' learning in the classes they teach?' The survey data clearly indicate that a large percentage (66.42) of college professors do not use the test type they believe to be best in both courses on which they provided data. The subjects' responses

were categorized into paradox groups from which the patterns of their test behavior can be described. Aside from the total agreement category (no existence of a paradox), the most popular pattern of test behavior were for faculty who use the test they believe to be best in their small class but not in their large class. This finding is important because it suggests that class size is an important factor in the decision of what test types faculty will use in their courses.

Approximately 20 percent of college faculty report that they use the same types of test in both their largest and smallest classes, but this test type is not their choice for best type. Once again this suggests that size of class and possibly other concerns are preventing them from using their preferred type of test.

Fifteen percent of the professors use one type of test in their largest class, employ a different type in their smallest class, but indicate a third type of test as best! (This is the Total Disagreement category.) This suggests that professors tailor evaluation procedures to fit their perceived needs, i.e. test fitting the course material, acceptable time required for grading.

As expected, professors tend to use the test type they prefer in their small class but not in their large one. The relatively low percentage of professors in the large class agreement category suggests that the logistics of a large class differ from those of a small class and those differences manifest themselves in qualitatively different faculty reactions. For example, in small classes, faculty tended to use essay questions while in the largest class they tend to

employ multiple choice tests. Complicating the picture, however, is the fact that many large classes are at the freshman/sophomore level while many of the small classes are upper division or possibly graduate level courses. As indicated in the interviews, decisions to use certain types of tests appear to be based primarily on four types of concerns: class size, the fit of the class material to the test, the faculty members' awareness of other people, and the relationship of tests with grades.

Some interesting differences exist in the data concerned with the relationship of how faculty test students compared with their beliefs about test types. Few differences exist between the percentages of faculty whose survey responses place them into the total agreement category or the total disagreement category. However, sizable differences exist in the large class agreement category, class agreement category, and to a lesser degree the small class agreement category.

Very few (2.78 percent) assistant professors fell in the large class agreement category while 13.91 percent for associate and 16.06 percent of the full professors fell there. This difference suggests that assistant professors generally do not use the type of test they believe to be best in their largest classes relative to associate and full professors. Assistant professors (25 percent) are more likely to find a test other than the one they consider best but that works for them and use it in both their large and small classes (class agreement) than their colleagues at the associate (13 percent)

and full (14 percent) professor level. Another way of interpreting this situation is that when designing the tests to be used in a course, assistant professors appear to be the professors most affected by large classes.

A second breakdown of the data was made by school size. The size of the school seems to have an impact on the subjects' testing behaviors since some differences exist in each of the paradox categories. The differences form an easily discernable pattern. Generally speaking, the faculty from small schools use tests they believe best more often than faculty from medium or large size schools. As expected, small school faculty fall in the total disagreement category least often, the large class agreement category most often, and the small class agreement category least often. All these characteristics combined, suggest that academic life at small colleges may be different than at the larger schools. This assertion is supported by the claim of several telephone interviewees that their decision to use certain types of tests was based on class size, job expectations, and the test's fit with the course material. It seems reasonable to assume that class size will differ somewhat in different size schools, and that expectations of faculty, in terms of number of courses taught, emphasis on teaching or research, etcetera, might differ along the lines of school size also. The job characteristics could play a major role in the decision of what test type(s) get used. The idea is that competition for time in terms of job expectations may be perceived as more profound in

larger schools than in smaller ones. For example, professors at small colleges and universities may be expected to engage in teaching almost exclusively, while teachers in the large schools are expected to engage in several activities outside the classroom, such that time becomes a less expendable commodity. Being expeditious with one's testing procedures may represent an acute awareness of time. The subject representing the small school category interviewed during the third phase of the data-gathering process mentioned time as a constraint when describing his experience of testing the fewest number of times of the three personal interviews. Before the interview began he described his role at this school as being primarily responsible for teaching.

The third breakdown of the paradox data was based on the subjects' academic areas. An interesting trend was noted in these data. The responses from subjects in colleges of education very closely approximated the respondents from the small schools. For example, education faculty had the highest group percentage in the total agreement category and the lowest percentage in the total disagreement and small class agreement categories. The implications of this similarity are discussed later.

Approximately 60 percent of the biology faculty's responses fell into the class agreement and small class agreement categories. Combining this with the relatively low percentages in the remaining categories, (especially large class agreement) it appears that biology faculty find it the most difficult to use tests they believe to be best.

When logistical constraints are least, i.e. when class size is down, they use their preferred types to a greater degree than any other academic areas in the study. This situation suggests that academic area has an impact on the types of tests used. This idea was corroborated by several subjects during the telephone and personal interviews. They noted that they chose their tests to a large extent based on how the test fits the material in the class.

The responses from psychology professors were placed almost evenly in all five of the categories. Their highest percentage was in the small class agreement, while their lowest percentage was in large class agreement. This indicates class size as a consideration, but the degree to which their responses fell into other categories suggest that other concerns exist too in their decisions to use the tests they choose. It should be noted that psychology is the only academic area where a higher percentage of responses exists in the total disagreement category than in the total agreement category.

The responses of faculty representing the category 'other' aligns them fairly closely with the respondents from psychology. An exception to this is that the 'other' category has a relatively high percentage of responses (38.16) in the total agreement category. Even with this being the case, the 'other' category still has a moderate percentage of faculty in the remaining categories. Based on the relatively large percentage of faculty from this group falling into the total agreement category and the percent in the other agreement categories, it appears that this group of professors

employs a wide variety of test types. Since this category represents numerous academic areas, interpreting such results must be done carefully.

Professors' Decisions to Use a Test Type

The fourth question motivating this study was, ' If some college professors do not use the type(s) of tests they believe to be the best method(s) for assessing students' learning, why not?' In formulating an answer to this question, it became obvious that professors' reasoning for not using their choice of best test type was the same as their reasoning for choosing a particular test type (Question 5: How do college professors decide what types of tests to use in the classes they teach?). For this reason, both questions will be considered together. The answer to these questions is based primarily on data gathered by the telephone interviews and to a lesser extent during personal interviews.

The data presented in Chapter 4 described the decision making processes of twenty respondents to the question of how they decide what types of test to use in their largest and smallest courses. From these data a category system was developed which provided a crude flowchart of the variables considered during the decision making process. The three variables--time, others, and fit--reoccurred throughout the interviews. In large classes, the most salient variable considered was the subjects' perceptions of time: whether they felt

constrained by time. The category of time included any factor which suggested that subjects were aware of elements which competed for their time (number of classes taught and other job requirements). For example, class size was included in the category others early in the reduction process. Later it was decided that the class size was really a component of a time constraint since many of the respondents noted the inordinate amount of time needed to grade the tests of a class with a large enrollment. It was not the people involved as much as the effort needed to score their tests. The remaining two categories, fit and others, did not play as important a role in the decision for the large class.

The responses for the small class indicated that, in most cases, class size was not a major factor in making the decision of which test type to use. Consequently, subjects made their decision based on their assessment of how a particular test fit the material in the course, or on various considerations of other people. The degree of fit had several components including level of students (freshman, senior, graduate), type of course (survey or skill building), amount of information to be covered. Numerous telephone interviewees noted these considerations.

The third category, others, had several characteristics. In many cases the subjects indicated they chose a certain test type because of the effects it would have on their students, because of administrators' expectations of them, and because of expectations of other faculty members.

As noted earlier, there appears to be a difference between how professors chose test types for the largest and smallest classes. Part of the explanation for this is obvious, the difference in the number of students in each class brings to bear different logistical concerns on the part of the professors. These logistical concerns are typically articulated as time constraints. When the time constraints do not exist, faculty are free to base their decisions on other considerations.

The Experience of Testing Students

The final question entertained in this study was, 'How do college professors describe their experience of the testing process?' Personal interviews of three subjects were conducted to begin an answer to this question. Subjects described a teaching/learning process which had at its heart testing students. They also described numerous things they were aware of when testing students. Interestingly, much of the teaching/learning process occurs either before or after the actual testing of the students. However, all subjects interviewed described this same process. The four themes which emerged against this (teaching/learning) ground attested to the existential nature of being a teacher. The themes--an awareness of time, others, their tests' fit with the material and objectives of the course, and the relationship of testing with grading--make sense given that professors operate in the world of their profession, a profession whose mission is to work for and with

people. To a large extent, the worth of professors is measured by how successful they are with other people. Therefore, not being aware of the various themes would be detrimental to one on the job. The degree to which subjects are aware of some of the themes seems to be affected by the job expectations placed on them. For example, the theme of time was not as figural for the subject from the small school as for the other professors. That same subject described his main job duty as teaching his courses. Engaging in research and university and community service were not expected. Prior to the interview, the subject described his time constraint as not having the time to engage in research. This supports the assertion that the experience of time as a constraint may be a "structural invariant" of the experience of testing students. The particular way this experience manifests could be affected by the situation one is in.

Implications of the Study

The various implications of the results of this study are embedded in the following discussion of the pervasiveness of multiple choice tests in our culture. Since multiple choice tests have been shown to be the type of test used most often, even though essay tests are considered to be the best type of test by the greatest number of faculty, it appears that professors choose their tests for reasons other than they believe it is best. Making this situation

worse is the fact that the multiple choice tests designed by professors and used in our colleges and universities tend to be poorly constructed, measuring only the lowest levels of Bloom's Taxonomy of Educational Objectives. Research indicates that the multiple choice test questions developed by professional test developers also tend to measure mostly factual information.

As described earlier (see Chapter 4), there appears to be a burgeoning market for computer test bank software that is being exploited by book publishers. If this new technological advance has the same degree of effect on our testing practices as previous advancements (the origination of blackboards, metal pens, and the development of a machine which could score thousands of answer sheets a day), it seems almost inevitable that multiple choice tests will become the only test type used to assess students in large college classrooms. The changes that have already taken place in our education system as a result of the advent of the "objective" tests have been marked. The consequences of these past events are that we now have in the United States two generations of students raised in a society where using multiple choice tests are status quo.

The following is a brief description illustrating how pervasive testing has become in our society. More importantly, it demonstrates how common multiple choice tests have become. A typical five-year-old boy begins his kindergarten career by taking a readiness test (which is given in a multiple choice format). At the end of first, third, fifth, eighth, and twelfth grades he must take

standardized achievement tests which are almost exclusively multiple choice tests. The twelfth grade test is a state minimum competency test. To go to college the child must perform reasonably well on a college entrance exam (i.e. Scholastic Aptitude Test), another multiple choice test. He chooses to major in the College of Education. To be admitted, he must take part of a professional teachers examination; once again, made up mostly of multiple choice test questions. While in college he takes the usual freshman/sophomore courses which range in size from 20 to 600 students. As seen in this study, he will likely get multiple choice tests in those classes. During the senior year, he must take the second half of the teachers' exam. He chooses to enter graduate school in Educational Psychology, but before being accepted, he must take some form of standardized multiple choice examination. And finally, after twenty years of school, he graduates with a Doctor of Philosophy Degree in Education with an emphasis in Psychology. Before he can practice his profession, he must pass a licensing exam, which is likely to be another multiple choice test. This example merely indicates how commonly students are tested, and how many important decisions may be made based upon their performance on a multiple choice test.

The ironic part about the current state of affairs is that many of the test experts believe that multiple choice tests have the potential to be equal to or better than any other type test in terms of their ability to measure a variety of levels of learning.

Unfortunately, however, most college professors never take a course in test construction. And even those who do find it quite difficult to write rigorous multiple choice test questions.

A possible explanation for the differences in assistant professors' testing practices and those of associate and full professors is that assistant professors are more idealistic in their beliefs about testing. They hold beliefs about testing which exist aside from any practical considerations. For example, numerous assistant professors considered essay tests the best type of test across all situations. To these assistant professors, a best test type existed. Conversely, many associate and full professors implied that even though they believed certain types of tests were better than others, they believed that the circumstances surrounding a particular class had an effect on not only their decision about which type of test to use, but also, which type was best. Assistant professors tend to hold beliefs about tests which exist in a purely academic context while associate and full professors tend to view tests in a more practical context. This difference suggests that there is a developmental change which occurs as one gains experience as a college professor.

Class size is cited most frequently by the subjects of this study as the reason they use objective tests. As professors are forced to deal with the various time constraints imposed by large classes, the students' learning is likely to be affected. Since some research suggests that different types of tests yield different types of

behavior from students and that this often translates into a difference in the type and degree of learning taking place, it is important for collge professors to exploit all types of tests for the benefit of their students. If professors continue to rely on only one type of test, the cost will be measured in terms of the type of learning which takes place. The effect of using one test type along with having a narrow focus in one's testing agenda is what Fredrikson (1984) called "the real test bias."

Suggestions for Future Research

The strongest attribute of the design of this study (using three different research methodologies to evaluate the validity of the data collected) also left some questions unanswered. For example, both Lazzara's technique and the phenomenological technique were adapted for this study. In doing so, the desired results from each were limited to the scope of this study. Consequently, neither technique was exploited to the degree that would be needed to answer some of the questions that need answering. For example, Lazzara's technique was developed to gather data which could be used to develop an expert system. Finding out how professors decide what types of tests to use in their classes could be empirically validated by conducting a study using only his technique.

Similarly, to uncover the structural invariants of the experience of testing college students one would need to use the

philosophy/technique of experimental phenomenology with a larger sample than what was used in this study. Even though themes did emerge during the data reduction phases of the study, the sample was not inclusive enough to provide the researcher with enough clues as to all the actual experiences, or the charateristics of those which did appear to truly describe the experience.

Limitations

The results of this study must be considered in light of the limitations of its design. This project has four known limitations: using a survey instrument developed by the researcher to gather data which allowed no room for qualifying answers, telephone interviews aimed at obtaining a very specific set of data, and the inherent limitations of using a dialogic method of inquiry with only three subjects. The study also was limited by the degree to which the researcher was able to collect and effectively thematize interview data into accurate units of meaning.

The use of three different research methodologies provided numerous benefits to the study, but it was necessary to adhere to different assumptions, unique to each methodology. In at least one case, a compromise was made that would not have been necessary if only one methodology had been used. When it was found that the survey data indicated that some professors' large classes had as few as six students and some of the smallest classes had 100 or more

students, the decision was made to let the data exist as received. This was done in an effort to allow the data to retain the respondents' intended meaning. Had the survey been the sole research methodology used, adjustments and reorganizations of the data would have been possible.

Despite these limitations, important information was explored in ways never before attempted on this topic. The limitations described would be more damaging to the study if only one type of procedure had been used. The three methods were chosen with an awareness of their particular shortcomings and the expectation of using the special view of the testing phenomenon each of them offered as a method of triangulating the data gathered. Therefore, the results of this study are believed to shed light on the college classroom testing phenomenon.

LIST OF REFERENCES

LIST OF REFERENCES

Ayres, L. (1909). Laggards in our schools. New York: Russell Sage.

Bailey, S. K. & Mosher, E. K. (1968). ESEA: The Office of Education administers a law. Syracuse, NY: Syracuse University Press.

Balch, J. (1964). The influence of the evaluating instrument on students' learning. American Educational Research Journal, 1, 169-182.

Barron's Educational Series, Inc. (1986). Barron's profiles of American colleges (15th ed.). Woodbury, New York: Author.

Black, P. J. (1968). University examinations. Physics Education. 3. 93-99.

Bloom, B. F. (Ed.). (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I, cognitive domain. New York: David McKay.

Bollenbacher, J. (1976). The testing scene: Chaos and controversy. In Testing in the public interest. Proceedings of the 1976 ETS Invitational Conference. Princeton, NJ: Educational Testing Service.

Bruning, J. L., & Kintz, B. L. (1987). Computational Handbook of Statistics. Glenview, IL: Scott, Foresman and Company.

Buros, O. K. (1933). Educational, psychological and personality tests of 1933 and 1934.

_____. (1938). The nineteen thirty-eight mental measurements yearbook. Highland Park, NJ: Gryphon Press.

Caldwell, O. W. & Courtis, S. A. (1923). Then and now in education. Yonkers: World Book.

Chapman, P. D. (1980). Schools as sorters: Lewis M. Terman and the intelligence testing movement, 1890-1930. Doctoral Dissertation. Stanford University (University Microfilms No. 80-11,615).

Chase, C. I. & Wakefield, L. M. (1984). Testing and grading: Faculty practices and opinions. Bloomington, IN: Indiana University, Bureau of Evaluative Studies and Testing. (ERIC Document Reproduction Service No. ED 256 196)

Colaizzi, P. F. (1978). Psychological research as the phenomenologist views it. Existential Phenomenological Alternatives for Psychology. Oxford University Press, New York.

Committee on Testing, Learning, and Grading of the Undergraduate Council. (1978). Survey of faculty testing in introductory courses. University of Tennessee, Knoxville. (Multilithed.)

Coulter, J.(1979). The social construction of mind. Totowa, New Jersey: Rowman and Littlefield.

Cross, T. L. & Huck, K. C. (in press). Successful tutoring program as the result of team effort. Tennessee Teacher.

Cureton, L. W. (1971). The history of grading practices. Measurement in Education. 4. 1-8.

Dapkus, M. (1982). A phenomenological analysis of the experience of time. (Doctoral dissertation, The University of Tennessee, Knoxville). Dissertation Abstracts International, 44, 302B.

Debra P. v. Turlington, 564 F. Supp. 177 (M.D. Fla, 1983).

Debra P. v. Turlington, 83 F. Supp. 3326 (1984).

Douglass, H. R. & Talmadge, M. (1934) How university students prepare for new type examinations. School and Society, 39, 381-320.

Dubois, P. (1966). A test dominated society: China, 1115 B.C.-1905 A.D. In A. Anastasi(Ed.), Testing problems in perspective (pp.29-36). Washington, DC: American Council on Education.

Eisner, E. W. (1988). The primacy of experience and the politics of method. Educational Researcher. June/July, 15-20.

Evans, E. D., Dodson, P. K., & Bailey, D. T. (1981). Assessment of psychology instructors perceptions and use of textbook test-item manuals for measuring student achievement. Teaching of Psychology. 8(2), 88-90.

Felker, D. B. & Dapra, R. A. (1975). Effects of question type and question placement on problem-solving from prose materials. Journal of Educational Psychology, 67, 380-384.

Frederiksen, N. (1984). The real test bias: influences of testing on teaching and learning. American Psychologist, 39(3), 193-202.

Gould, S. (1981). The mismeasure of man. New York: Norton.

Gronlund, N. E. (1985). Measurement and evaluation in teaching (5th ed.). New York: Macmillan Publishing Company.

Halpin, Glennelle & Halpin, Gerald. (1982). Experimental investigation of the effects of study and testing on student learning, retention, and ratings of instruction. Journal of Educational Psychology, 74, 32-38.

Haney, W. (1984). Testing reasoning and reasoning about testing. Review of Educational Research. 54(4). 597-654.

Harre, R. & Secord, P. F. (1972). The explanation of social behavior. Oxford: Blackwell.

Heidegger, M. (1962). Being and time. New York: Harper and Row. (Original work published 1927)

Houts, P. (Ed.). (1977). The myth of measurability. New York: Hart.

Ihde, D. (1979). Experimental phenomenology. New York: G. P. Putnam's Sons.

Kruger, D. (1981). An introduction to phenomenological psychology. Pittsburgh: Duquesne University Press.

Lazzara, P. J. (1985). Foundations for a method for knowledge analysis. Unpublished doctoral dissertation, University of Tennessee, Knoxville.

Lunneborg, P. (1977). College grades: What do professors intend to communicate? Seattle: Assessment Center, University of Washington.

MacGillivray, W. A. (1985). Embodiment and ambiguity: A phenomenological investigation of the experience of the lived-body. Unpublished doctoral dissertation, University of Tennessee, Knoxville.

Madaus, G., Kellaghan, T., Rakow, E. A., & King, D. J. (1979). The sensitivity of measures of school effectiveness. Harvard Educational Review. 49. 207-230.

May, R. (1969). Existential Psychology. New York: Random House.

McNames, N. (1977). Summary of responses to testing practices questionnaire. Kansas City: KCRCHE.

Merleau-Ponty, M. (1962). The phenomenology of perception. New York: Humanities Press.

Meyer, G. (1934). An experimental study of the old and new types of examinations. I. The effect of the examination set on memory. Journal of Educational Psychology. 25. 641-661.

_____. (1935). An experimental study of the old and new types of examinations. II. Methods of study. Journal of Educational Psychology, 26, 30-40.

_____. (1936). The effect on recall of the examination set in classroom situations. Journal of Educational Psychology, 27, 81-99.

Milton, O., Pollio, H. R., & Eison, J. A. (1986). Making sense of college grades. San Francisco: Jossey-Bass Publishers.

Nairn, N. & Associates. (1980). The reign of ETS: the corporation that makes up minds. Washington, DC: Learning Research Project.

Nicolson, F. W. (1917). Standardizing the marking system. Educational Review. 54. 225-237.

Nolt, J. E. (1984). Informal logic: Possible worlds and imagination. McGraw-Hill, New York.

Parker, S. C. (1912). A textbook in the history of modern elementary education. New York: Ginn and Company.

Popham, J. (1978a). The case for criterion-referenced measurement. Educational Researcher. 7, 6-10.

Popham, J. (1978b). Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall.

Popham, J. & Husek, T. (1969). Implications of criterion referenced measurement. Journal of Educational Measurement. 6. 1-9.

Romanyshyn, R. D. (1977). Phenomenology and psychoanalysis: Contributions of Merleau-Ponty. Psychoanalytic Review, 64(2), 211-223.

Rugg, H. O. (1915). Teacher's marks and marking systems. Educational Administration and Supervision. 1. 117-142.

Scannell, D. P. & Tracy, D. B. (1975). Testing and measurement in the classroom. Boston: Houghton Mifflin Company.

Semb, G. & Spencer, R. (1976). Beyond the level of recall: An analysis of higher-order educational tasks. In L. E. Fraley and E. A. Vargas (Eds.), Third National Conference on Behavior Research and Technology in Higher Education. Gainesville, University of Florida.

Smallwood, M. L. (1935). An historical study of examinations and grading systems in early American universities. Cambridge: Harvard University Press.

Spencer, R. E. & Stallings, W. (1970). The improvement of college level student achievement through changes in classroom examination procedures. Final Report Program, #6-1174. USOE. Urbana: University of Illinois.

Starch, D. & Elliot, E. C. (1912). Reliability of grading high school work in English. School Review. 20. 442-457.

Starch, D. & Elliot, E. C. (1913). Reliability of grading high school work in mathematics. School Review. 21. 254-259.

Terry, P. W. (1934). How students study for three types of objective tests. Educational Research, 27, 333-343.

Thayer, V. T. (1928). The passing of the recitation. New York: Heath.

Thomas, L. & Augstein, S. (1970). An experimental approach to learning from written material. Uxbridge, England: Center for the Study of Human Learning, Brunel University.

Trani, E. (1979). Helping students become more sophisticated consumers of their own education. FIPSE. Lincoln: University of Nebraska.

US Bureau of Census. (1975). Historical statistics of the United States, colonial times to 1970. (Vols 1&2). Washington, DC: United States Government Publications Office.

Valle, R. S. & King, M. (Eds.). (1978). Existential-phenomenological alternatives for psychology. New York: Oxford University Press.

Von Wright, G. H. (1963). Norm and action. The Humanities Press, New York.

Wert, J. E. (1937). Twin examination assumptions. Journal of Higher Education. 8, 136-140.

Winograd, T. (1980). What does it mean to understand language? Cognitive Science. 4. pp. 209-241.

APPENDIXES

Appendix A

The Survey Instrument

Based on the last completed academic term in which you taught courses, please provide the requested information for your class with the largest enrollment and your class with the smallest enrollment in the spaces provided below.

	Largest Class		**Smallest Class**
1) How many students were enrolled in this class?	_____		_____
2) What percentage of the students' grades were tied to their test performance?	_____		_____
3) Put a check beside the type(s) of tests used in this class.	_____	Essay Exam	_____
	_____	Fill-in-the-blank	_____
	_____	Multiple Choice	_____
	_____	Oral Exam	_____
	_____	Short Answer	_____
4) If you checked more than one type of test in question three, which type accounts for the bulk of the students' test scores? .....→	_____	True False	_____
	_____	Other (specify)	_____
	_____		_____

5) In your opinion, what type of test allows for the most accurate assessment of students' knowledge and skills? (See note on back of accompanying letter.) Circle <u>one</u> option.

Essay Exam **Fill-in-the-blank** **Multiple Choice** **Oral Exam**
Short Answer **True False** **Other(specify)** _____



NO POSTAGE
NECESSARY
IF MAILED
IN THE
UNITED STATES

BUSINESS REPLY MAIL

FIRST CLASS PERMIT NO 477 KNOXVILLE TENNESSEE

POSTAGE WILL BE PAID BY ADDRESSEE

Learning Research Center
The University of Tennessee
1819 Andy Holt Avenue
Knoxville, Tennessee 37916-9989



Appendix B

Letter to Subjects Accompanying Survey



Learning
Research
Center

Learning Research Center
The University of Tennessee
1819 Andy Holt Avenue
Knoxville, Tennessee 37996

Dear Professor,

We are gathering data which we hope will provide insights into the testing practices of college professors in the United States. Enclosed with this letter is a card containing five questions. Please take the time to answer these questions and drop the card in the mail. The card was carefully designed to take the least amount of time and effort to complete. In addition, the postage has already been paid.

Since a high "response rate" is needed for the results of this survey to be considered valid, it is critical that you return the requested information. At no point will your name or any other identifying characteristic be divulged, so please be as candid with your responses as possible.

We are very appreciative of your willingness to help with this research project. If you have any questions about what is being requested, or if you are interested in the results of the study, please write or call.

Thank you for your help!

Sincerely,

Tracy L. Cross

Tracy L. Cross
Ph.D. Candidate

W. Lee Humphreys
Director, Learning Research Center
Professor of Religious Studies

The University of Tennessee, Knoxville • 1819 Andy Holt Avenue, Knoxville, TN 37996-4350 • (615) 974-2459

When thinking about question five, limit your response to the type of test that you believe provides the most accurate assessment of students' knowledge and skills. Do not compromise your beliefs by tying them directly to logistical concerns. Make your decision based solely on your beliefs about the properties of the types of tests.

Appendix C

Letter to Testing Service Directors

THE UNIVERSITY OF TENNESSEE
KNOXVILLE

2705 Sanderson Road
Knoxville, Tennessee 37921
(615) 524-4227

College of Education

Department of Educational and Counseling Psychology

Dear Test Service Director,

I am gathering data to be used on my doctoral dissertation that deals with the extent to which teacher-made objective tests are used in the college level classroom. I am trying to discern the pervasiveness of these type of tests by documenting the numbers of machine scorable answer sheets that have been disseminated to faculty at selected schools over a forty year period. I am collecting the data to develop a brief history of the numbers of answer sheets sold at a target population of schools. This will provide me with a conservative estimate of both the overall number of objective tests used at the college level as well as a discernable trend of the objective testing practices of college professors.

If it is possible from your records, I am requesting that you send me a chronology of the number of answer sheets sold or given to faculty that were not used for standardized tests. The number of answer sheets cited need not be concerned with the numbers returned to be scored by your service. At this time I am only interested in the total number of answer sheets sold or given to faculty. It is very important to my research that I document these numbers, so please mail me a copy of the numbers of answer sheets disseminated by year. I am particularly looking for a format similar to this one:

1941 - 11, 313
1942 - 14, 645
1943 - 15, 873
.
.
1985 - 32, 854

If you can provide additional information related to this request, for example, if any department tends to use more answer sheets than others, I would appreciate that information as well. I will be combining this information with other data in an effort to paint a picture of the testing practices of college professors.

* Since the selection of schools for this portion of the data gathering procedures is very small and select, it is critical to my research that you return the requested information. If there is no testing service at your school, please forward this letter to the person who could best respond to my request.

I am very appreciative of your willingness to help me with my research. If you have any questions about what is being requested, or are interested in obtaining the results of my study, please call or write.

Thank you for your help !

Sincerely,

Tracy Lee Cross

Tracy Lee Cross
(Ph.D. Candidate)

108 Claxton / Knoxville, Tennessee 37996-3400 / (615) 974-5131

Appendix D

Letter to Software Companies

THE UNIVERSITY OF TENNESSEE
KNOXVILLE

2705 Sanderson Road
Knoxville, TN 37921
(615) 524-4227



College of Education

Department of Educational and Counseling Psychology

Dear Software Distributor,

I am collecting data to be used for my doctoral dissertation that deals with the extent to which teacher-made objective tests are used in the college classroom. In order to complete my research procedures, I need to obtain a list of your company's sales records of computer software that is used to generate and/or store objective test questions. A description of the software packages you sell would be helpful also. I am particularly looking for a format that looks like this:

Year		Software sold
1972	-	87
1973	-	108
	-	
	-	
1985	-	673

Since the selection of software companies for this portion of the data gathering procedures is very small and select, it is **critical** to my research that you return the requested information. Please note: ***Company anonymity will be respected at all times.*** Any additional information you want to provide concerning sales of computer item banking software will be appreciated.

I am very appreciative of your willingness to help me with my research. If you have any questions about what is being requested, or are interested in obtaining the results of my study, please write or call.

Thank you for your help!

Sincerely,

Tracy Lee Cross

108 Claxton / Knoxville, Tennessee 37996-3400 / (615) 974-5131

Appendix E
Telephone Interview

Subject #134 S 30 P5

I: Well, last year about this time I believe you received a postcard-sized survey with questions related to this. At the top of the survey it was as you were asked several questions. The first question was how many students were enrolled in your largest class and you had indicated 35. Now this was a class which you taught that previous academic term.

Q Right.

I: Would you have any idea which class this might be?

Q Well, yea, if it was 35 it would be my general psychology class. Those classes are usually from 35 to 50.

I: Ok. What. . .and you were asked what percentage of the students' grades were tied to their test performance and you said 100%.

Q That's right.

I: And that the type of test you used was a multiple choice.

Q That's right. That would be my general psych course, because I don't use multiple choice in anything but that. Yeah, that's clearly the general psych course.

I: Ok. And the final question was in your opinion, what type of test allows for the most accurate assessment of students' knowledge and skills and you had indicated essay exam. Ok. Now the question I would like to ask you is this, I would like you to think back to when you were considering what type of test you wanted to use for this class and I was wondering if you could describe how you came to your decision.

Q Yeah, it's very simple if you have 35 to 50 students in a class, it is virtually impossible to do essay exams and that is why I always use multiple choice in the introductory class.

I: Alright.

Q Curiously enough, I don't use multiple choice in any other course and the reason for that is those courses usually have fewer students.

I: I see.

Q And when I can use essay exams, that is when I can manage to survive the paperwork, then that's what I always use.

I: Alright, so that's a primary consideration.

Q Oh yes! You see if I had 15 or 20 students in my general psychology class I would probably use essay exam, but when you get up to 35 to 50, that's just impossible.

I: Alright. Well, was there anything else which entered into the decision?

Q Well, I suppose in a way, because I've always given multiple choice exams in my general psychology course, I have a good back log of information on the validity and reliability of the various test items on each of the tests.

I: I see.

Q And I do an item analysis of the multiple choice questions after every time I give it and that enables me then to either change some questions or to throw them out or to salvage them, if they are salvageable, so that over the years with that kind of data, you can actually build a very, very good test.

I: Ok, and this is a class which you have taught for some time?

Q Oh, indeed I taught that class for longer probably than you are old.

I: Oh, goodness. Ok, so you feel that the multiple choice test which you have come to use for this class is a good indicator.

Q I will guarantee it.

I: Alright.

Q As a matter of fact, the students, my multiple choice exams are kind of notorious around the college, because I can actually get them to do things other than memorize facts, with multiple choice tests you can put in analogies or you can put in things that really do require them to relate things to each other and to see similarities and differences, so my tests over the years have gone away from just being memorizing facts to trying to get them to really think.

I: I see.

Q And over the years I have become pretty confident that I can get them to do that.

I: Well that's good. Now that was with your largest class.

Q Right.

I: Ok. Now we asked the same questions concerning your smallest class and you had indicated that there were ten students in it and that there were. .

.approximately 75% of the students' grades were tied to their test performance.

Q Right.

I: And the test of choice was essay exam.

Q Right.

I: Alright, would you happen to remember which class this was?

Q Yes. If it had ten in it and was last year, right?

I: Yes.

Q Ok. If it had ten in it then that was my learning class.

I: Alright. Well, I would like to ask you the same questions.

Q Sure.

I: I would like you to think back to when you were considereing what type of test you wanted to use for this particular class and if you could describe how you came to your decision.

Q Well, in a way I have already answered that, but I will put it in this context. Clearly in my opinion, you are more likely to be able to get students to think and to relate things to each other and to see similarities and differences if they are writing about what they know.

I: Yes sir.

Q In a class of ten I could manage that with no problem at all. So, there is really never any question about what kind of exams they would have.

I: Alright. So you feel like the essay exam really tests more of what they actually know?

Q Yes, and not just the factual information that they know, but their understanding of what they know about the facts that they know, if that makes any sense.

I: Ok. I see in terms of being able to synthesize that data.

Q Yes. In other words, you can get them to write a summary of something in an essay which will give them the chance to, synthesize is a good word, to analyze and synthesize what they have learned about a particular thing, rather than just spitting out one isolated fact about it and to me that is much

more important.

I: Alright, was there anthing else you would like to add to that?

Q No, except that in addition to getting an understanding of what they know, I think writing essays gives them practice in writing and think that is a very important thing to know how to do, too.

I: Yes sir.

Q And I always stress clarity and conciseness of expression when they write my exams. So my students learn pretty quickly not be bring a shovel to class. (laughter) If you get my drift.

I: Yes sir, I sure do.

Q So I think it is very good experience for them to practice in writing about what they know clearly and concisely. So, I think that is an added advantage of the essay exam as opposed to a multiple choice test.

I: Alright. Well, is there anything elss you would like to add concerning. . .conversation?

Q No, I think that just about covers it. You asked some good questions.

I: Well, great. I appreciate the time that you spent.

Q You're welcome.

Appendix F
Full Arguments of Telephone Interviews

060 M 13P2 (Total Agreement) Large Class

1. Using essays allows me to write questions that require the students to apply the material as well as integrate material from different sections of the book or from different lectures. (Rule)

2. By using essay exams, the students are required to learn and understand the material better than if they were taking an objective exam. (Belief)

3. The way I do it is simpler than creating other test types (drawing a subset of questions for the test from a list of questions previously given to students.). (Belief)

4. By drawing test questions from a previous handout, the students must learn all the material. (Belief)

5. I'm not a very good multiple choice question writer. (Belief)

6. I don't want to test simply for memorization of material. (Want)

7. It is difficult to write a multiple choice test that will allow me to determine how well they learned the material and whether they can apply it. (Rule)

Summary: Essay exams require the students to learn the information, not just memorize it.

Small Class

1. I used essay to try to tap into not only how well they learned the material but also whether or not they apply the material. (Intention)

2. Other test types would not allow me to measure how well they learned the material and whether they can apply it. (Belief)

3. It is more important to learn the material than it is to just memorize it for the test. (Belief)

Summary: Essay exams show me how well students learn the material.

266 M 20E3 (Total Agreement) Large Class

1. When you're working with people who intend to become administrators, they need to be able to communicate effectively both in speaking and writing. (Belief)

2. By using essay exams students can learn to express themselves effectively in writing. (Fact)

3. Essay exams give people more flexibility to express themselves. (Rule)

4. With essay exams students are more able to use their knowledge and be unique in expressing themselves. (Rule)

Summary: Essay exams allow students to express themselves and develop their writing skills.

Small Class

1. I have a philosophical orientation toward developing and continuing an ongoing dialogue with these graduate students. (Belief)

2. A dialogue can be maintained through the use of an exam that requires the students to write. (Rule)

3. I wanted to provide them with the types of questions that will encourage the interaction and determine not only what writing skills they have but what conceptual abilities they have and that's why I made the decision I did. (Want)

Summary: Essay exams promote a better understanding of what the students are learning.

206 M 18E2 (Large Class Agreement) Large Class

1. In the past I used a variety of questions on each test, but students complained about the essay questions so I didn't use them anymore. (Fact)

2. When there are complaints from students, the instructor should make changes in his test. (Belief)

3. If you're able to construct adequate multiple choice items you can test the individuals understanding at a number of different levels. (Belief)

4. If a multiple choice exam is constructed adequately, it will do a good job of testing the students. (Belief)

Summary: Multiple choice tests should be used since students don't like essay questions and since the information can be covered as well with multiple choice.

Small Class

1. In the past I used a variety of questions on each test, but students complained about the essay questions, so I didn't use them anymore.

2. When there are complaints from students, the instructor should make changes in his test.

3. If you're able to construct adequate multiple choice items, you can test the individual's understanding at a number of different levels. (Belief)

4. Good multiple choice tests require skill to develop. (Belief)

Summary: Students do not like taking essay exams and the information can be covered as well with objective exams.

202 S 3005 (Total Agreement) Large Class

1. By gradually shifting from one type of test to another over time, I came up with a test that fits my needs for this class. (Belief)

2. I think if you were to have essay exams you would be forced because of time constraints to have only a few of them and the students need more feedback than that. (Rule)

4. Tests should be used which provide enough feedback for students.

5. We've got to come up with a common language and multiple choice lends itself very nicely to that. (Rule)

6. Exams other than multiple choice do not eliminate the problems associated with semantic and hermaneutic interpretations of language.

Summary: Multiple choice tests provide a common language between teachers and students and provide ample feedback to students.

Small Class

1. The multiple choice exams were by and large things for them to get a little feedback early on. (Rule)

2. The multiple choice tests could be graded more quickly and the students can get a feel for where they stand before the end of the course. (Fact)

Summary: Multiple choice exams can be graded quickly and can provide feedback to the students.

481 L 09O2 (Total Disagreement) Large Class

1. Students want their grades back immediately. (Fact)

2. Objective exams can be graded quickly and returned to the students faster. (Fact)

3. In large classes if you spend all your time grading papers, you will be criticized by your peers for spending all your time doing it. (Belief)

4. If one spends all his time grading papers rather than doing other scholarly things, then he will be criticized by his peers, consequently, one should not choose tests which will elicit the criticism of colleagues. (Belief)

5. Multiple choice exams are necessary. (Belief)

6. When teaching a large class one must use multiple choice exams in order to have time for research. (Rule)

Summary: Objective exams allow the instructor to get them graded quickly and back to the students, thus allowing more time for valued activities.

Small Class

1. In that kind of a class, it is because that is what a professional should be doing. (writing). (Belief)

2. Students who are going to be professionals should be competent in writing. (Belief)

3. Essay exams allow students to demonstrate their abilities in writing. (Fact)

4. It also goes back to how I was treated as a student and it's one I always thought was a proper one. (Custom)

Summary: Essay tests allow students to improve their writing skills.

306 M 21P5 (Total Disagreement) Large Class

1. One reason for the decision was that I wanted to cover a greater breadth than depth of material on the test. (Want)

2. Tests cover differing degrees of depth. (Rule)

3. Multiple choice tests are better for covering the breadth of the information. (Belief)

4. I wanted them to be more familiar with terms; with the basic concepts. (Want)

5. Multiple choice tests best cover terms and basic concepts. (Belief)

7. Another reason for objective tests is the number of people in the class. (Rule)

8. A lot of time is needed to grade tests other than multiple choice for a large number of students. (Rule)

Summary: Decision is based on size of the class as well as a desire to cover a lot of information.

Small Class

1. I wanted to probe for depth of knowledge. (Want)

2. I feel that essays are a better instrument to probe deeply into a knowledge base and thinking. (Belief)

3. In the case of essay exams the student is required to go deep within his knowledge base to come up with the answer. (Rule)

Summary: I wanted to probe for depth of knowledge and essay and oral tests were best for that.

181 L 03P1 (Total Disagreement) Large Class

1. I have to use multiple choice when the class is large. (Rule)

2. When the class is large, you do not have time to grade essay tests. (Rule)

3. The essay test just seems to be more reasonable and the students can take a stab at things and qualify their answer. (Belief)

4. With essay tests, you can see whether they can clearly express their thoughts. (Rule)

5. Essay tests come closest to seeing if students have mastered the concepts. (Belief)

6. Multiple choice tests cannot measure whether students can express their thoughts clearly. (Belief)

7. Objective tests cannot measure whether students have mastered the concepts. (Belief)

Summary: Even though essay tests can benefit the students in a variety of ways, multiple choice must be used because of the time required to grade essays.

Small Class

1. I had to see if they could memorize large amounts of material in a short time for what I call the skills test. (Rule)

2. Tests should match the type of information being tested. (Rule)

3. One can ask more questions on an objective exam and therefore cover more material. (Rule)

4. Another type of test covered conceptual material and I felt that short answer and essay tests were a more appropriate way of testing that kind of information. (Belief)

Summary: The type of information tested determines which type of test will be used.

307 M 21P5 (Total Disagreement) Large Class

1. I wanted to attempt to sample adequately from the amount of material which has been covered. (Want)

2. Supply test questions do not allow the degree of sampling of covered material that recognize items do. (Rule)

3. It is possible to have a large number of items on an objective test because it can be graded quickly. (Fact)

4. Another factor is time. (Rule)

Summary: Objective tests are used in order to cover a large amount of material in a reasonable amount of time.

Small Class

1. I wanted to attempt to sample adequately from the amount of material which has been covered. (Want)

2. Supply test questions do not allow the degree of sampling of covered material that recognize items do. (Rule)

3. It is possible to have a large number of items on an objective test because it can be graded quickly. (Fact)

4. Another factor is time. (Rule)

Summary: Objective tests are used in order to cover a large amount of material in a reasonable amount of time.

135 S 30E5 (Large Class Agreement) Large Class

1. Just for the ease of grading, I use multiple choice. (Rule)

2. Other types of exams require too much time for grading. (Belief)

3. With multiple choice I think that I can evaluate students as well as with any other type of test, because I write my own test. (Belief)

Summary: Multiple choice exams are used because of ease of grading.

Small Class

1. That course is designed to be more of a creative course. (Belief)

2. Giving them an essay test is the only way to look at their creativity and thir ability to act beyond the material. (Rule)

3. Giving essay tests allows the students to express themselves more and go beyond the material. (Rule)

4. Other types of tests cannot measure creativity effectively. (Rule)

Summary: Essay exams allow for better evaluation of the students' learning.

206 S 3005 (Large Class Agreement) Large Class

1. My decision depends on the size of the class.

2. The practical utility different tests offer is more important than tests' measurement characteristics.

3. In a class with 45 or 46 students I use multiple choice tests.

4. Tests other than multiple choice should be used in classes with few students.

Summary: I select my tests on the basis of size of the class, level of the class, and the kinds of things the students are doing.

Small Class

1. My decision is based on the size of the class. (Fact)

2. In a graduate class with 14, 15 or 16 students, I use essay tests. (Fact)

3. My decision is also based on the kinds of things they're doing in class.

Summary: I select my tests on the basis of size of the class, level of the class, and the kinds of things the students are doing.

012 S 22B1 (Large Class Agreement) Large Class

1. I use a variety of question types because some people don't do very well on some question forms and this way we would be leveling things out. (Rule)

2. It is important for students to have more than one type of question on tests so they can compensate if they are not successful on a particular type of question. (Belief)

3. If I gave a strict essay test it would take me forever to grade the thing. (Rule)

4. It gives them a chance to write a little bit and think a little bit and at the same time gives me a chance to grade the thing in a reasonable amount of time. (Rule)

Summary: Students benefit from more than one type of question since everyone is better at one type of test question than another and less time is involved in grading tests that are not strictly essay.

Small Class

1. I would much rather have them tell me a little bit about what's behind things.

2. Fill in the blank and essay tests are used so students have to come up with the answer from what they've learned, not just pick an answer.

3. The major reason for taking that approach is to try to get them to think critically about what they might read in magazines and newspapers.

4. When they have to know the information well enough to come up with their own answer, they are more likely to think critically about what they read.

Summary: Supply tests teach the students to think critically.

080 M 14P2 Large Class

1. I use tests which allow me to make sure the students are getting the content of their reading. (Belief)

2. The test also is to have them pull together their ideas and to synthesize what they have read. (Belief)

Summary: The test allows better evaluation of the students' learning.

Small Class

1. I decided what types of tests to use based on the size of the class.

2. You have to think of a matter of time; what you have to do in addition to all the other things we do as a professor.

3. The choice of which tests to use cannot be made in a vacuum; time limitations must play a part in the decision.

4. In a small class you can get more intense.

5. Since there are fewer time constraints associated with the small class, I can use the types of test I believe to be best.

Summary: My decision was based on time limitations.

011 L 01P3 (Small Class Agreement) Large Class

1. The size of the class determines largely the types and kinds of tests I use. (Law)

2. Decisions about tests must be based on logistical concerns. (Rule)

3. Anytime a class gets over 50 or 60 students, I almost will always go to a multiple choice exam. (Rule)

4. Tests other than multiple choice should be used in classes of less than 50 students. (Rule)

5. It is too time consuming to grade essay tests for a class of over 50 students. (Rule)

Summary: Since essay tests are too time consuming to grade for a large class, the size of the class determines the type of tests used.

Small Class

1. It gives them an opportunity to review when I give them questions like that. (select a subset of questions from a list of questions previously distributed.) (Rule)

2. I wanted them to demonstrate to me that they have worked with the material on their own. (Rule)

3. Essay tests allow the students to demonstrate their ability to integrate and synthesize the material. (Rule)

Summary: Essay tests require the students to demonstrate that they have worked with the material.

190 M 17E2 (Small Class Agreement)

Large Class

1. The reason we use multiple choice tests is because there is so much content. (Rule)

2. Multiple choice tests can be graded more quickly, therefore you can ask more questions and cover more material. (Fact)

3. The multiple choice just seems to be the most appropriate way to test them on the information that we give them in the course. (Belief)

4. Tests should fit with the type and amount of information covered in a class. (Belief)

Summary: Multiple choice exams are the most appropriate way to test certain types of information.

Small Class

1. The multiple choice just seems to be the most appropriate way to test them on the information we give them in the course. (Belief)

2. Tests should fit with the type and amount of information covered in a class. (Rule)

3. One can ask more questions on objective tests because it doesn't require as much time to grade and more questions means you can cover more material. (Rule)

Summary: The multiple choice exams are the most appropriate way to test certain types of information.

134 S 30P5 (Small Class Agreement) Large Class

1. If you have 35 to 50 students in a class, it is virtually impossible to do essay exams and that is why I always use multiple choice in the introductory class. (Rule)

2. With a large number of students, essay exams require too much time for grading. (Rule)

3. By using multiple choice questions, it is possible for me to check the validity and reliability of the test questions. (Belief)

4. Over the years you can build a test that you know is a good test. (Belief)

5. My tests do a good job since over the years I have developed a valid and reliable set of test questions. (Belief)

6. I can actually get students to do things other than memorize facts. (Rule)

Summary: The objective test is the practical thing to use in a large class and, over the years, one can build a good objective test.

Small Class

1. You are more likely to be able to get students to think and to relate things to each other and to see similarities and differences if they are writing about what they know. (Fact)

2. When students are required to write answers to essay test questions they are more likely to better understand the material. (Belief)

3. I think writing essays gives them practice in writing and I think that is a very important thing to know how to do. (Belief)

Summary: Essay tests are used to make the students think and relate things to each other.

002 L 01P1 (Small Class Agreement) Large Class

1. Even though I do not like multiple choice tests, my decision to use them was based on the practical realities of large class.

2. I chose this test based on the degree of work needed to grade them.

3. I use this test type because students bitch less when given multiple choice tests.

Summary: Multiple choice tests are most easily graded.

Small Class

1. Since small classes are manageable, the students don't have to get ready in such a short period of time to take their tests. (Belief)

2. In small classes, students can come in and take tests on their own time.

3. Testing should not take time away from instruction.

4. Since I'm principally a researcher I use multiple choice tests because they are better, easier to grade than other types of tests.

5. The testing process should be as time efficient as possible so as not to take time away from other duties.

Summary: The testing process should be as time efficient as possible so as not to take time away from other duties.

305 M 21P5 (Class Agreement) Large Class

1. My procedure has been to give objective exams for a long time. (Custom)

2. I use machines to score my answer sheets because of convenience and that's 99% of the decision. (Rule)

3. I have a theory of faith in my objective exams because I write them myself. (Belief)

4. Because I write my own multiple choice exams I have customized them to the point where they give me confidence in their ability to measure what I desire with the added convenience of being able to be graded quickly. (Belief)

   Summary: I have confidence in my multiple choice tests' ability to measure what I desire and they are the most convenient test type.

Small Class

1. I believe in expediency.

2. I have already got the multiple choice tests ready and I typically use that. (Custom)

3. It is more convenient to use a test already prepared than to write a new one. (Rule)

   Summary: Multiple choice exams are used for expediency.

484 L 09O2 (Class Agreement) Large Class

1. Decision is based solely on the size of the class. (Rule)

2. Certain class sizes dictate what types of tests to use. (Belief)

3. My decision in choosing a multiple choice format and a short answer format is based upon my inability at about a number above 40 to be able to grade examinations and turn them back to the students within a reasonable period of time. (Rule)

4. I base my decision on what types of tests to use on the importance of getting feedback on their performance on a test. (Belief)

5. I don't think I lose a great deal in multiple choice exams. (Belief)

6. Multiple choice tests are not the best type of tests to use. (Belief)

7. Because I spend a lot more time writing the multiple choice test questions than I do essay test questions, I think they do a good job of testing the students. (Belief)

Summary: Class size dictates the type of test to use and, although not the best test type, multiple choice tests can be effective and give the best and most timely feedback to students.

Small Class

1. I felt that there were certain areas of the material that required the student to really prove to me that they could come up with the right words in order to describe some phenomena in genetics. (Belief)

2. Short answer questions require students to prove their knowledge. (Belief)

3. Multiple choice was also used because of the size of the class--larger than my limit for entirely short answer exams. (Fact)

Summary: While class size is a limiting factor, it is necessary to use supply tests to be sure students are learning the right things.

215 S 31E2 (Class Agreement) Large Class

1. I could cover the material better with essay tests. (Belief)

2. Due to the types of questions asked, it was not possible to use objective exams. (Rule)

3. Objective tests are limited to asking only certain types of questions.

Summary: Essay exams fit the needs of this class better than objective exams.

Small Class

1. I wanted my students to learn a lot of skills in that class. (Want)

2. Objective tests were inappropriate for the course goals. (Rule)

3. The skills that needed to be learned could not be tested using multiple choice tests. (Rule)

4. Objective exams would not aid in improving writing skills. (Belief)

5. The short answer were more appropriate. (Rule)

Summary: Objective exams would not help to achieve the goals of the course.

065 S 26E2 (Class Agreement) Large Class

1. The logical thing to do with a large class is to have very objective questions that can be scored through the computer. (Rule)

2. Due to time constraints, when the class is large, objective tests need to be used. (Rule)

3. With short answer responses I can see how students' attitudes have changed toward the material. (Rule)

4. It is not possible to perceive the students' attitudes toward the class when the tests are computer-scored. (Belief)

Summary: When the class is large, objective tests should be used, conversely, when the class is small, short answer tests should be used since they allow the instructor to see attitude changes toward the material.

Small Class

1. With the types of test I use, I can find out more about how they use things or how they feel about things or how their attitudes have changed toward the topic. (Rule)

2. I get more unique input than I could giving a different type of test. (Belief)

3. Recognition tests could not provide me this kind of information. (Belief)

4. When students are writing their own answers, I can get an idea of how they feel about the material. (Rule)

Summary: I can better evaluate student learning with supply tests than recognition tests.

Appendix G
Personal Interview

Small School Subject #1
Interviewer=I, Subject=S

I: What I would like you to do is think of a test or a testing situation you have been involved with that stands out for whatever reason, and describe it to me.

S: Well, I guess what's on my mind right now is the General Psych course, because that is what I'm teaching this summer. Of course here we teach General Psych every semester; there might be one exception to the schedule. And I guess that's been on my mind because it's during the summer and I'm always tinkering and trying to figure out the best way to work that course in general. And of course part of that is how do I evaluate what they're getting out of it. And the objectives that I have of the course in a liberal arts curriculum is in part to teach them how to think a little better and to immerse them in some of the critical issues in our culture, as well as to teach them some psychology and some of the critical issues in psychology, and how we go about addressing those issues as a complicated member of society should think, should try and do, and decide which of those you can test, or testing is measurable in any way--of course they're all measurable I guess--but measurable in a way that you can also do. But one of the things that I remember let's say, as an incident, is making multiple choice tests. There was a section where we were studying brain and the sensory projection areas, and you know in the sensory projection areas of the brain there are parts of the body which are very sensitive and have large projection area, and there are parts of the body, like the small of the back, which are not very sensitive, and have a small area. Now the test could talk about how differences in size in these areas varies. I used a. . .I believe I probably used a multiple choice item for projection and said if you had an elephant, and had to lok at the sensory projection area, which part of the sensory projection area would be larger, that which came from the tip of his trunk, or that which came from the back of his leg? Clearly we were trying to stress that the trunk is very sensitive compared to the back of the leg, and when I was going over the test afterwards, one of the students was very irritated. He said "You never said anything about elephants in class." And it was very funny, because what struck me with this total, well I think part of it was unwillingness to deal in the abstract issues, to draw from the sheer fact, and I guess that is one of the things that I struggled with, to try to get them to learn some factual data, something that's in the textbook, look at the implications and think about it, and engage in some sort of higher level mental operation, with the information. Now that tells you, of course, that I use multiple choice.
What I do . . . General Psych, of course, typically a freshman is taking it as a distribution requirement. Not quite as bad as a required course that everybody has to take, because it's an elective distribution, but it's still a requirement, and you have the problem of the haze that comes over their eyes and the glaze, when they're in one of these courses that you're supposed to take. But what I try and do, I test differently in my other

a final--to make them, what I'm trying to do is structure it so they have to read that textbook on a regular basis. I'm trying to build them the habit of reading chapters every night, and test them every so, every two chapters, so they develop some study habits. Of course they groan about that, and about the number of tests a semester, but the other side of that is that many of them have done well enough in high school without having to develop any discipline, and they start sinking at the college level. Our school is a pretty busy, demanding place, and my idea is to develop consistent discipline, so if they sink, it's because they're not doing the work, it's not because they don't have the chance to learn how to study. So I give them eight tests a semester, so that's a pretty high level. I wouldn't think about doing that in an upper division course. And, what I have settled into, can become hazardous, is I usually, each test is usually half multiple choice and the other half is essay. What I try and do in the multiple choice is see if they are reading and comprehending the text. I use multiple choice items and I use a lot of the structured items because I like them. And I ask them to also, if they're having difficulty with the workbook, because I want to, and I use those multiple choice to see if they're reading the textbook and if they're comprehending the textbook. And I do, I have an undergraduate test course/psychological measures course which covers all of that, but I do try to have some simple, factual-oriented questions, some more difficult thought questions that they have to figure things out, some generally complicated ones and some analogies. I try to vary the difficulty and also the type that they have. Typically I want twenty multiple choice and then twenty points of essay, usually two ten-point essays. But, again, what I want them to do, what I tell them I'm going to do with the multiple choice is that I'm going to see if they are reading the textbook and comprehending it, and usually I can find out if the publisher's items identify what I want them to get out of the text. But I'll always do a little shuffling as I go through because I teach the course every semester, I usually keep records on the mean scores on the multiple choice and the difficulty level of the first couple of semesters when I use the textbook, to get some idea about the reliability. And then the essays, I used to just ask them essays, but I have discovered again that the incoming freshmen, they haven't learned to focus on the bigger issues, so what I do for each chapter is I give them ahead of time, I sit down and figure out three salient issues in that chapter that I also know something about, and I give them ahead of time for each chapter three potential essay questions that I can ask. Some of them can be answered from the text, but most of them can only be really answered if they've also been in class. So I tell them, these essay questions are not going to be tied to the text like the multiple choice. The text will help you, but also lecture will help you, and make it clear, to get them to come to class, that if you're going to do well on the essays you're probably going to have to be in class, or at least get somebody to come. So I give them three ahead of time, and then on the test, give them--I cover a test every two chapters--I have one essay question from one chapter and one from the other. But they've already had six essay questions ahead of time, which could be one of those six. My idea again is that I'm going to use the essay questions to give them a little focus, and then get their discipline for studying to get those essay questions, get together in a group to figure out

what I want or what I mean, work out your answers ahead of time, you know, do all that stuff, and you come to class, to take the test, you could already have it written out, but you can't hand in your pre-written, you have to write it out, but I'm telling you, it's a sure thing, if you really want to do well on the course, well, that's the only way to do it. Most of them don't do that, of course. So, I also try and make the workbook, Intro Psych Text Workbook, is optional. It focuses on the basics, and it's training in test-taking, because the workbook asks fill-in-the-blank with multiple choice and other stuff, but it's good training in using that information, and if you don't do well on the first multiple choice test, that means you're not tracking the text close enough, so you're going to have to use that book for the training. But if you're good readers, and tracking well, you don't have to worry about it. So, on the first test, there'll be twenty multiple choice items and two essay questions, from that group of six possible, and what I do then, of course, is have a machine score the multiple choice, and then I hand-score the essays. And then the next day in class--and I really feel pretty strongly about giving the test back the next day, certainly not much later--we'll go over the test, and I feel like we really need to go over the test. I'll put the distribution on the board, the class distribution, so they can see what the class distribution was. And then we'll go over the multiple choice. Usually I read the questions and ask somebody to tell me what the answer is, and it takes a while and I encourage them, if you've got any questions or you disagree or you don't understand, say so, because that's a chance to find out about the item. And, so we go over each one, and people don't ask many questions, which after the first test or two they find out that I really am interested in whether they understand and we can use the review of the multiple choice items as another learning experience. And also, I introduce them to item analysis if there's real controversy, people really disagree what they think the question said, and so on, the way to resolve this issue if you're a psychologist is to back up and find out how many A students passed this item, how many B students, C students, D students, and so on, and then look at that distribution, and I'll have Doug, here do that, and he'll bring it in, and maybe only ten students, maybe only five out of thirty passed it, so see, that's a different measure of hard and ambiguous. But if the A students got it, the B students, some of them got it, none of the C students got it, I'd say well gee, I guess it was just hard. So anyway, I talk to them about how do you decide if an item is ambiguous or just hard? And they like that, and they go and bug the other teachers about it, because, as you mentioned, not many teachers know much about testing, at least you pick up (laugh) enough that you can win them over and then they go bug some of the other . . "Will you let us do an item analysis on this question?" " What's that?" "Well, Dr. ___" (laugh)

But, we go over the multiple choice, and then I collect the multiple choice, I don't like to leave those with them, because they say, "Well, we want to study the tests for the final."I say "No, I want you to study the text." I get back all the multiple choice. Then we go over the essay, and what I usually do for the essay is, each one is worth ten points. A fourth of it is forty points on each test, twenty points are multiple choice and twenty are essay, so each essay is worth ten points, and I ask a student who got a ten to read his answer. And. . .if they're willing. And usually one, or sometimes I'll ask

two students to read, and then tell them what I saw in that answer what made it a ten, so they get some idea of why they were right or wrong. They mentioned this, they mentioned that, they mentioned that, and someone says "Well, I mentioned all those, why didn't I get a 10?" And I explain the difference between a nine and a ten is, and I explain about clarity, coherence, and so on, which I say, one thing about writing essay questions, it's not just that you've got it all there, but you've gotta be convincing. You know, you've got to be clear in preparing it. So, and there's a bunch of factors, I get arguments, you know. They come up after class, "Well, I've got everything you said," and then I have to try and convince them. They eventually get the idea though, that in the essay questions, it's not just that it's all there, but are they convincing. But, since it is a freshman course, and I don't really have the time, I tell them I'm not going to grade them on spelling, punctuation, and grammar, if I can understand what they are saying. That takes some pressure off of the students, and you get some wonderful spelling. My favorite is schizophrenia, and it came out literally skitso frantic, which really has a kind of ring to it, you know, because if you take it literally, schizophrenic. But that makes them a little uneasy, I'm not grading them on spelling and grammar, but I am grading them on whether it's coherent and convincing, because one can have the grammar and spelling there, and all of the facts there, but still not get a ten. But, or another has miserable spelling and grammar and gets a ten. And that bugs some of them, but I keep hammering that this course I don't have time to do the English grammarian. But they accept the other side of that, "Well, your grading is subjective," well, that's right, there's bound to be some of that in there, but you get a nine if you've got everything there and you get a ten if you got it there in good form. But that's basically what I do with that, and out of those eight semester tests, I tell them how I'll drop their one lowest test. Again, giving them the impression of course that you're allowed to blow it, with that many tests, at least once. So at the end of the semester, I will drop their one lowest test score and. . .

I: I got the message that--tell me if this is accurate--that your students are willing to argue with you more about your markings on their subjective part of the test, essays, than they are about the objective--is that right?

S: Actually, in terms of sheer frequency, they are more likely to ask questions about the multiple choice. They don't--I guess what happens with the essay is that they will come up after class--they don't argue in groups--but they will come up afterwards and say "I don't understand." Their frequency--I don't know. My impression is that in the group they'll ask questions and challenge the multiple choice more. But after it's that kind of a--after class they're going to come up it's not about the multiple choice, it's about their essays. It becomes a private thing. I think if you just look at sheer frequency, it is probably more on the multiple choice.

I: When they talk to you about the objective test items, do you have a feel for the types of arguments they are making, or the questions they are making?

S: Um, usually my impression is that they didn't read the textbook closely and

they picked up on one part but not seeing the big picture. The simplest thing that comes to mind is that we have a chapter on stress mamgement the following is one way to reduce stress, and it lists four ways to reduce stress, and the correct answer is "all of the above." And they picked up on one and they will say, "Well, it says right here in the text that so-and-so reduces . . ." and I'll say "Yes, it does, but if you refer to this, this and this, this does too." So very often it is incomplete knowledge. But . . . I did do an item analysis on one where it was an interpretation problem and the question could be interpreted either way, unless they read the text very closely. Item analyses of it was supported by the better students agreeing with me, but one of them pulled out the workbook and the workbook had a reversal in it, and the workbook said something else, and not everybody used the workbook, but I figured if the workbook, and it was an error in the workbook, so in that case, I gave them credit for either one. There's not too many--it's not too many it's a matter of logic or so on, because over the years you get pretty good at picking out or writing questions which are not too ambiguous, though usually it's a matter they are hard enough--they'll say they are tricky--but what they mean is they didn't do a close reading of the chapter. That may be the. . .

(Side B)

I: Their answer, three things you said needed to be, and they didn't agree with the score ?

S: Yeah, yeah. That's typically the problem, is that they think they've said it and I don't, and I have to go back and show them where they didn't convince me, that what they said could be interpreted other ways. And I change some of them, I uh, don't change too many, but I do, because sometimes it is my problem. What I try to do when I score the essay questions is I try to have numbers, numbers instead of a general idea, and I figure out what I think is an answer, what the essentials are, and I go through and read three or four of theirs to get an idea what they're doing, then I go back and really start grading the papers. I do find--and then I'll do all of one essay question, I'll find that I try and look at their essay questions in an internal sense. But usually by the time I've scored ten or twelve of their essays on one question, I have developed a pretty good subjective sense of whether they've got it or not. But that's sometimes hard to convince them of. I say, "Well, you know, actually I scored a bunch of these. You seem to have it all, but you just gave me, you know, it wasn't like you really understood, and that's why you didn't get that extra point." They don't like that. But I try and explain--I guess that's in part phenomenological--that after I've scored quite a few of them, I get some sense of what is there and what's not, and that's ultimately part of the subjectivity. But again, I'll also notice that when I do give in to a student, one or two, I get a rash more of others try it on the next test. The word gets out that he can, he will do it on occasion, sometimes, so there's another part of me that thinks if I change this one, then I'll have to argue with a whole bunch more, and I don't want to do that, so I try and toe the line where I can, in part because I believe that they didn't write a good essay. It's probably not as ethical, but it does happen.

I: Normally, while the students are taking their test, are you in the class with them?

S: Yes.

I: What do you do during that time?

S: Uh, usually what I do is I take my textbook with me, well first of all I hand them out, and then I walk the class--I, usually it's in the large class I have them spread out a little bit, and I usually tell them jokes, I've been reading somewhere, probably Howard Pollio, or something that you know, a little tension reduction. And sometimes they groan and sometimes they laugh, but that's--if I don't tell them a joke now, they say "Where's the joke?" And even if they groan they still want their joke, and then, after I've passed out the test, I walk around to make sure everybody's got one and not two, and just kind of monitor and just kind of proctor them in a sense. I can also tell if they might have crib notes or something like that, though I haven't seen too much of that kind of cheating, I do see sometimes people looking at other people's papers, so I'll call down the classroom and have them spread out.

I: Does any particular--does anything come to mind, standing out for you, during the test administration, do you remember one incident at all over others?

S: Well, uh . . I usually walk the room then I get out my text and I start taking a look at the next chapter, and when they start bringing in their essay questions, I have them turn in the multiple choice here, their answer sheet here, and their essay questions here, and I will start shuffling through their essay questions and looking at them, just trying to get a feel for what they are doing, and then periodically look up at the class, and then I usually walk around at least another time, just to kind of make my presence felt. I guess I'm very aware of their nonverbal behavior, and I do tell them if they have any questions, raise their hand, and I'll come to them, rather than having them disturb the others. I guess the thing that distresses me, and is hard to deal with, is when I see a student really trying to look on another student's paper, and I don't like it, but also I realize that a lot of times when you're taking a test you kind of wander around, so then I start keeping an eye on the student, to see if they just are um--just a kind of vague behavior on their part, or whether they're doing it, and, then, if they are, uh, cheating. Since you can't be absolutely sure what's going on, I feel like I should confront them but what I usually do the first time, I see them, and then I double check, and if it does look like they're trying to see what's on somebody else's paper, I will stop them after class and say, "It . . appears to me that you were you know, looking at somebody else's paper and I'm not sure if that is what you're doing, but it looked like it," and see what they have to say. Of course, a lot of the time the majority of them will deny it. "Since I'm not sure, and I do feel strongly about it, I think next time I'll place you." Usually I let them sit where they want to except spread

out a little more. And, uh . . .yeah.

I: Are you comfortable with that?

S: Sure. And that usually takes care of it. But, I guess I remember that because I find it upsetting. I don't like to confront somebody. I'm not a confronter sort, I'm not a real hardass. I prefer to assume the best and work for that, and yet I feel very strongly about it, so I . . .

I: Does one particular incident come to mind? Is that what you're telling about?

S: Yeah, I guess what I'm thinking of is one of our students and an international student. It's interesting, um, because our international students tend to be very, very competitive, and they also tend to be more--I may be wrong, but I think they tend to be more a kind of person who's learned to work with at bureaucracy and the system, more likely to push the system to its limits. They will argue more about grades, they will argue more about the syllabus, they will argue more about everything. You know the kind of image I have--I've done some international work with the Peace Corps--working with volunteers as a trainer--and this is it--bureaucracies in many countries are malleable to someone who knows how to go in and shake it. And in general, international students know how to do that better. This one student was that kind of student, whom I had to spend a fair amount of time with explaining the system and he was also looking at other people's papers, and I confronted him about it. He was so hurt that I would ever think such a thing, and so on and so on, you know, just, and I must admit I didn't believe him. I thought he was putting his best face on, I think you've got one of those types. There was enough history there that it was uncomfortable. I didn't particularly like him, which made it hard too. . . but at the other time I was convinced he was . . . so, and I did ask him to sit in another place. He was willing, but not without going through a lot of dramatic denials and postulations or whatever . . . and subsequently I discovered that he was failing in other courses, so I don't feel bad about it. If things like that are upsetting to me, and most, you know, are, it's their future and I feel very strongly, and can't understand how anybody would do that, and you know they're cheating themselves more than they are the system, and all this stuff, but uh, it takes me a while to verbalize my energy to stop them, because that part of their spirit, they are still kids, go through everything we can to help them, so that they will succeed. I'm one of the the more tenderminded sorts on the faculty here.
But when I'm sure, I have to do something. Yeah, I remember, that's the kind of thing I remember in class. Most of the time testing goes fairly smoothly, and I rather enjoy going over a test, because I find that as we go over the items, as far as things go, I can use them as a chance to talk a little more about the material.

I: I got the impression early on that you use your tests for a lot more than just assessment. You mentioned going over, just a couple of other things that indicate you try to get the students to read systematically, and read

thoroughly and carefully and so it sounds like you have an image of testing other than or in addition to evaluation.

S: Yeah, I do view it as a structure to enhance the whole process, and I talk to them about that, I don't talk to them about testing and studying and everything until after the first test, because it goes in one ear and out the other, until they've had a test. But then I do go back and after the first test say you know how well you did will apply to the final grade. If you're on the right track, and if you had trouble, or didn't do as well as you wanted, here are some things you might think about, and I go into a bit about how to read the chapter, you know, skim it before you come to class, at least; I talk to them about skimming and what it is and what it isn't. Most people don't understand, they find it hard to do. Then you hear--and I take kind of a cognitive approach--you skim it, you know you look at the outline summary, the major headings then you've got the big picture. Try and don't . . don't read the textbook in detail, don't try to copy from it, don't try doing that the first time. Then listen to me, and I will continue filling in the details. I see my role in class is to try to put the material in context, both related to their life world and to larger issues. After . . I say then, and only then, you've skimmed it and listened to me, do you go back and read the textbook slowly, carefully, and in detail, because if you try and do it slowly, carefully and in detail the first time, you go through a lot of detail that's not going to stick. So I give them that kind of spiel, then I talk about you know, doing a little bit every night is better than waiting until the end, and that I'm going to ask them to do something that's bizarre and cruel, to read it a second time. I talk to them about repetition and explain to them that I can't go in and read that chapter and lecture to them after one reading, even though I've been teaching for twenty years, because I can't get it all. So I talk about how they start--my approach, anyway, is to start with a framework, try to get a framework, and then fill in the facts, and explain to them that those facts are going to stick if we have some schema or whatever term, I use a spiderweb metaphor, or something like that, until we get to the chapter on cognition and then say "That's what I was talking about." But, and I've thought about relating it to their life and discussing it, how can they study with friends. Some of them have this bizarre, it's almost like cheating if you work out your essay questions ahead of time and talk about it in groups. So I do that after the test too. But I guess part of it is it's an introductory course, and I figure it's one of those you ought to spend more time structuring and better cognition and better structuring and to get them to use the normal process. And we get you know usually about a third of the students in there are upperclassmen probably bored to tears or putting them to sleep or whatever. But I explain that it's important that uh, it's a freshman level course. This is what we're about. And testing is part of the structure. And so I go to an upper division course, I don't do it that way. I use less multiple choice, more short answer and essay. In fact I, I teach Abnormal Psych, and that's a kind of a survey. I do use some multiple choice in that because it's a survey course and I want to make sure they're getting something from the detail. Other courses, I use short answer and essay. There is a size factor. If I was teaching at UT and had sections of a hundred, whatever, I wouldn't attempt essay questions. We set a limit,

formally at thirty in our Introductory Psych and we'll let in up to thirty-five if there's a special problem. With thirty I find I can do the multiple choice and essay, but it does make it rough. I have some trouble when I teach Abnormal, usually that gets to be pretty big, too, I might get thirty or forty in that, and it's hard to do that for the essays.

Second Reduction of the Personal Interviews

Subject #1 (Small School)

1-1 I am aware of class size when I decide what test type to use. If I had sections of a hundred, I wouldn't attempt essay questions. With thirty, I find I can do the multiple choice and essay, but it does make it rough.

1-2 I feel strongly about giving the test back the next day. The first test has 20 multiple choice items and two essay questions. I have a machine score the multiple choice and I hand-score the essays. I am aware of the need for quick feedback.

1-3 I am aware of the difficulty of making students engage in some sort of higher level mental operation; to look at the implications of factual data and think about it. In one class we had discussed the sensory projection areas of the brain and their size in relationship to the sensitivity of their respective body parts. The multiple choice item said "Which part of an elephant's sensory projection area would be larger, that which came from the tip of his trunk, or that which came from the back of his leg?" One of the students was very irritated and said, "You never said anything about elephants in class." What struck me was the total unwillingness to deal in the abstract issues, to draw from the sheer fact.

1-4 In the lower division courses, I test the students frequently. I am aware that the preparation for the tests helps to build good study habits.

1-5 I am aware of my responsibility to give the students a chance to learn how to study. Many of them have done well enough in high school without having to learn any discipline. Our school is a busy, demanding place and my idea is to instill a consistent discipline with frequent testing.

1-6 With incoming freshmen, I discovered that they haven't learned to focus on the bigger issues, so I give them a set of essay questions beforehand, one of which will be on the test. My idea is to use the essay questions to give them a little focus, and then use their discipline to figure out the essay questions in advance. I am very aware of their need for direction and how I can use the test to teach them discipline.

1-7 In a freshman course, I tell the students I'm not going to grade them on spelling, punctuation and grammar if I can understand what they are saying. I am aware of how much time that would take. I get some wonderful spellings because of that. My favorite is schizophrenia which came out skitso frantic.

1-8 I am aware of the subjectivity in my essay grading, but I am comfortable with it and the students accept it. You get a nine if you've got everything there and a ten if you got it there in good form.

1-9 I am aware of a difference in the students' attitude toward the multiple choice items and the essay. In the group, students will ask questions and challenge the multiple choice more. But the essay becomes a

private thing; students come up after class and say, "I don't understand."

1-10 Very often the reason a multiple choice question is missed is because they picked up on one part but are not seeing the big picture; incomplete knowledge. I am comfortable with my test items though, because over the years you get pretty good at writing questions which are not too ambiguous. Students say they are tricky, but what they mean is that they didn't do a close reading of the chapter.

1-11 When I'm grading essay questions, I'm aware of the subjectivity of it. When the student thinks they've said something and I don't, I have to show them where they didn't convince me. I change some of them, because sometimes it is my problem. But sometimes it's hard to convince them they're wrong. They don't like it.

1-12 I notice that when I do give in to a student, I get a rash of others who try it on the next test. I am very aware of the complications that arise from my acknowledgement of the subjectivity of my grading and subsequent willingness to compromise with the students. If I change one, then I will be inundated with students who want me to change their scores next time, and I don't want that. So I try to toe the line where I can.

1-13 On test day, I am aware of the tension the students feel. I try to reduce that by telling a joke. They have come to expect it.

1-14 I am very aware of the students' nonverbal behavior during test taking. I walk around the room, just to make my presence felt and I tell them to raise their hand if they have a question and I'll come to them, rather than disturbing the others.

1-15 Part of my awareness of their nonverbal behavior is focused on the possibility that they are cheating. It destresses me and is hard to deal with when I see a student trying to look at another student's paper. I realize that a lot of times during a test you kind of wander around, so I keep an eye on the student to see if it was just a vague behavior or actually cheating. I feel strongly about it.

1-16 I had to confront one student about cheating and it was upsetting. I can't understand how anybody would do that, and you know they're cheating themselves more than the system. It takes me a while to verbalize my energy to stop them. I think that, since they are kids, we go through everything we can to help them succeed.

1-17 I've been trying to figure out the best way to work one course in general, and it makes me aware of the beginning of the testing process--the design of the test. The course objectives are defined, so I try to decide which of them is measurable in a way that can be done.

1-18 When we go over the test, I am aware that the students need to know that I am really interested in whether they understand the material and

that I see going over the test as another learning experience. If there is controversy over an item, we discuss how to decide if an item is ambiguous or just hard. They like that and they go and bug the other teachers about it.

1-19 I am aware of testing as a method of building skills. I view it as a structure to enhance the whole process. After the first test, I talk to the students about how they did and what they can do to improve.

1-20 I am aware that certain test types are more appropriate for some courses. In a survey course, I use multiple choice because I want to make sure they're getting something from their reading. Other courses, I use short answer and essay.

1-21 I enjoy going over the test, because I can use the items to talk a little bit more about the material. I am aware that the test does not have to be the end of the learning process.

1-22 When I am creating a test, I am aware of what I want each item to do as far as evaluation goes. I use multiple choice to see if they are reading and comprehending the text. I try to vary the difficulty and the type of questions the tests have. I keep records on the mean scores on the multiple choice and the difficulty level to get some idea about the reliability.

Subject #2 (Medium School)

2-1 I am aware of the time needed to go over all the questions, explaining the answer.

2-2 I am aware of the time required for the type of test I plan to give when there are so many other things I need to get accomplished. I've thought of giving a shorter test, but I rarely do.

2-3 I am aware of the time it takes to build a test. It depends on whether you have to start from scratch. You can cut and paste and revise from older tests, and that isn't nearly as painful as starting from scratch.

2-4 The size of the class makes me aware of how I can use the test. If I had 55 or 60 people in there, I might give them the same kind of test, but I might not spend as much time going over it with the idea of teaching from it. What is the practicality in beating yourself to death by trying to be fair to the students?

2-5 I am aware of the difficulty in grading essay tests. I have little patience for reading nonsense and trying to piece together what someone's trying to say. I typically give objective tests.

2-6 I am aware of the need for good quality tests, but I don't have the time, or haven't taken the time, to write multiple choice questions for one class. That's a painful job, but I really need to do it.

2-7 I am very aware of class size when I think about testing. It's the number of students and the topic that dictate how I test. If we had 30 people in that graduate class, I'd be real angry, because I would feel forced to give the same kind of tests that I give. I want to make sure they know how to do certain things. But I'd be very unhappy about having to grade them. At some point, I'd say the heck with this, we're planning a multiple choice test.

2-8 I am aware of the differences in students' grasp of the material. The ones who could work the problems and master the concepts I thought ought to be rewarded with a really good grade.

2-9 I am aware of the significance of students' grades. I wasn't about to use a quantitative class to drum people out. On the other hand, there's a notion that it's sometimes pretty easy to get A's in graduate courses. I want it to be a little bit challenging.

2-10 I am aware that doing well on a tough test can be rewarding, but I don't want to punish people who didn't quite handle it all.

2-11 When I don't give students credit for a question, but they're close, I just tell them "if your grade in the course depends on you getting this question correct I'll give it to you." I am aware of the importance of grades and one question shouldn't make a grade difference. But if you

give enough tests, it's not likely that it will happen.

2-12 One of the reasons I am so aware of the quality of the test is because it is the major part of the grade. If a single student is on a borderline between one grade and another, I'll consider their class attendance and participation, but I rarely use any subjective kinds of evaluation in my courses. I like to use a lot of tests.

2-13 I am aware of the students' opinions of my tests. They liked taking two parts of the test in one night. It's a big occasion, but I think they liked it.

2-14 I am aware of the students' opinions. In my undergrad statistics class, they don't want concepts tests because we cover more topics in a wider, less detailed manner.

2-15 I was aware that the students who did well in both areas seemed to like the test.

2-16 The students made me aware of their feelings about the test by commenting that they really felt like they had learned something.

2-17 I am aware of the demands of test taking, so I always tried to create a nice testing environment where they wouldn't feel like I was trying to harm them in some way. They'd laugh about something even while they were taking the test. The environment was good.

2-18 I am aware of the subjectivity of essay tests. Students have said to me that they do better on essay tests and I think "because you get away with bullshit too often, and I'm not going to let you do that." I think overall they probably like objective tests better--they are easier to take, you don't have to generate or start your own response.

2-19 I like teaching about testing in part because I am aware of how much undergraduates like to complain about tests. I tell them, "you're going to know about reliability and validity and when you want to complain about a test now, you can use some words that'll intimidate the faculty members--value analysis, p-value, etc.."

2-20 I am aware of students' opinions of our tests. In our evaluations, students rate the item that says test evaluations are made in a fair and impartial manner below average at all levels, individually, department, college, university. You can just see the whole pattern, the profile.

2-21 I am aware of testing as a teacher effectiveness tool. Sometimes when you look at a test you've given, you get some clues to your instructional deficiencies. When certain items are missed by a sizable number of people, or even just a few, it's worthwhile to go over that topic again. One more shot at instruction. I think it's a real good use of tests--after evaluation, continue instruction based on the responses to the test.

2-22 I am aware that the test is not the end of the teaching process. It's easy to give tests back, but I think they need to be restated, rehashed. The students don't always understand that I want them to learn the material. They will say, "Why don't you just give me my grade?"

2-23 I am aware of the importance of being comfortable with testing as an evaluation of students' knowledge, but I am unhappy about the way I evaluate the students in my parapsychology class. Even though I don't like it, semester after semester, the distribution of scores is very similar, so I must be doing a fairly good job.

2-24 Sometimes there's not much of a deep discussion you can have about things in a course where there's a lot of remembering events, cases, etc. rather than some deep theoretical kind of thing. Going over the test is less useful then, but I do it anyway. It's just a habit.

2-25 I am very aware of the need for the content of the test to reflect the things you've been doing in instruction--topicwise. I may take a limited amount of material out of the reading assignment, but the primary content of the test will come out of things we actually discuss in class.

2-26 From the influence of my professors, I became aware of the importance of constructing good tests. I try to make sure the content is there and the grammatical sense--any other cue that might cause someone to get the item incorrect.

Subject #3 (Large School)

3-1 I am aware of the time required to take the test. The test was comprehensive and the students had three and a half to four hours to take it.

3-2 I am aware of the different time requirements each student has. Some people finish early and leave, while others are there even when the time is up and I end up staying another fiver or ten minutes.

3-3 Sometimes I wonder why I give essay exams when I am aware of the time required to grade them. I tell the students to expand and to write what they want and make sure they come to the point. Later I am sitting there with three days to get grades in and 20 stacks of essay exams. Then I wonder why I did this.

3-4 I am aware of the students' need to express themselves fully on an exam. Objective tests do not allow a student to express himself or get out as much information as he wants to. I feel very strongly that people ought to be creative about this. They may know some information my question doesn't quite get at, and being able to explain their rationale seems a healthy thing to me.

3-5 When I give a test, I am aware of the way it is based on the objectives I have given the students. The students weren't shocked; they knew what was coming in the test. I just asked for details.

3-6 In a testing situation, I have to be aware of the level of student. This was a 6000 level graduate class, a very self-selected group. I've never felt comfortable giving an objective test at that level.

3-7 Because of my awareness of the student level in my courses, I don't believe that undergraduate students have a lot of background and information that they can bring into a test. I use short essay to give them the opportunity to expand.

3-8 When grading tests, I am aware of the students' ability to write well and to the point and organize information in addition to having all the points. With some people, it's unorganized and you have to wrestle with it.

3-9 When I make out an essay exam, I am aware that what I am looking for is particular knowledge. I jot down the points I want and arbitrarily try to figure how much credit to give. If the student is edging on to a point or they incorporate it in another part of the test, I try to give them credit for it.

3-10 Even though I encourage students to tell me why they're thinking a certain way, I am aware that their answers can get out of hand. The answers get a little bit longer than you like. I've tried all kinds of ways to force them to be more concise, but I don't want to defeat my other

purpose. . .to expand.

3-11 I am aware of the testing environment. Our school has an honor code system. We don't have to be there or monitor a test. The students know where I am if they have a question during the test.

3-12 I am comfortable with giving tests at the students' leisure because of the strength of the honor code. I give them a sheet to sign that they pledge that they have neither given nor received help on the exam. And with the honor code, if they sign the sheet, that's alright. There are strict penalties, so I think it works. There will always be people who cheat no matter what.

3-13 I am very aware that tests should be built around stated objectives. You know what you want the students to do as a result of your teaching. I give detailed instructional objectives beforehand, then try very hard to build my tests around them. I don't like the concept of testing as punishment which it would be if there were unexpected questions.

3-14 I don't object to the tests. I am aware that I can use them as a teaching device. I like to come in and say, "Okay, I'll have a test today," and then go over which points they got and didn't get.

3-15 I go through a topic and I think I've done well and when the students get to the lab, the can't even get the thing turned on. So sometimes you have to use other ways to see if you're getting the point across. I like to use tests for that. If you give the students your objectives and you teach and they don't get them, you can't say "That's a bunch of dumb people." You have to say, "What did I do as a teacher that was wrong?"

3-16 I am aware of my reputation as an educator. I do not like the idea that a lot of good grades mean you're easy or not a good teacher, or that a lot of bad grades mean you're a good, tough teacher. The College of Education has gotten a bum rap on grade inflation, but I think it's because we take testing more seriously. I let the students know what I am going to hold them responsible for. As a result, they generally do well.

3-17 I love teaching, but the thing I have never liked about it from the first day is grades. I'd really be in seventh heaven if I could teach and interact and do my thing and then when grade time came, not have to do it. I am aware that they're necessary, but I don't give them lightly and I spend a lot of emotional energy on grading.

3-18 I don't like giving grades because I am aware that no matter how quantitative you try to be, there is always a subjective judgment. People consider grades important and I don't want to make mistakes. I don't want to deny a guy a grade because of three points.

3-19 I think it is my responsibility to hold the students accountable for what I said they would be. At the same time, colleagues say they give easy grades in Education. My retort is "Come in and sit with my class. See

how easy it is." I am aware that my primary responsibility is to the students.

3-20 When I'm working with a graduate class, I am aware of the quality of student at that level. They are self-selected, a dedicated group of people. I don't think at the graduate level we work on a normal curve. I think it's a very skewed curve. I don't believe in grading on a curve. If a class of 10 people all did terribly on a test, there are going to be 10 F's. If they deserve an A, that's what they're going to get.

VITA

Tracy Lee Cross was born in Knoxville, Tennessee on June 25, 1958. His formal academic training began in Knoxville at Pleasant Ridge Elementary School in 1964 and culminated in a Doctor of Philosophy degree from the University of Tennessee in 1988.

His current interests include the effects of classroom testing on learning and the psycho-social development of gifted and talented students. He assumed the position of Assistant Professor of Psychology in the Department of Behavioral Sciences at Louisiana Tech University, Ruston, Louisiana, in 1987.

Tracy is married to Jennifer Riedl Cross and has two sons--Ian Tracy and Keenan Patrick.