An investigation of the effects of missing data technique selection on student performance results

by

David James Putz

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Education

Major Professor: Richard P. Manatt

Iowa State University

Ames, Iowa

2000

UMI Number: 9962840

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Graduate College
Iowa State University

This is to certify that the doctoral dissertation of

David James Putz

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

Signature was redacted for privacy.

For the Graduate College

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I. INTRODUCTION

Berube (1994) contends that as historians look back over the 20th century, they will document three broad reform movements in American education. For the first half of the century, the Progressive Movement shaped education. It sought to educate the whole child, focusing on intellectual, social, and moral development. The desire was to produce knowledgeable, productive citizens (Parker & Parker, 1995). The second major reform movement was the Equity Movement. The civil rights movement of the 1950s and 1960s influenced the Equity Movement's focus of educating the poor (Berube, 1994). The United States is currently in the midst of the last broad education reform movement of the 20th century, the Excellence Movement.

The Excellence Movement was seen as having an economic impetus (Berube, 1994; Murphy, 1992). A strong bond was formed between business and education. Spurred on by matching concerns over America's declining economy (Berube, 1994) and the declining performance of school children (DiConti, 1996; Ravitch, 1995), both political and business leaders believed that "the skills and expertise of a country's work force are essential ingredients needed for greater economic productivity and national prosperity" (DiConti, 1996, p. 15).

During the Excellence Movement, a heightened level of attention has been paid to all matters connected to education. Though much attention is given to the plethora of reports that were released in the early 1980s, many states had enacted legislation and moved to improve education prior to the release of the reports (Kirst, 1988; Warren, 1990). This being known, the year 1983, when the National Commission on Excellence in Education released its report, *A Nation at Risk*, is often accepted as the start of the current reform movement (Kirst, 1988). The state-level activity geared to improving education reached an all time high in the 1980s (Center for Policy Research in Education [CPRE], 1989), and the reform reports of the 1980s helped support a sustained effort to improve American education (Murphy, 1990).

Large numbers of national and state reports were released in the 1980s with the intent of informing and directing education reform. As many as 16 national and 175 state-level reports were released for circulation by 1983 (Feir, 1995). This large number of reports and the actions taken to impact education have often been referred to as 'waves' (Manatt, 1993; Murphy, 1990; Parker & Parker, 1995; Plank & Ginsberg, 1990). Because of the number of efforts, it is not possible to definitively outline what happened when, with some feeling that the reform efforts have run parallel with each other (Baker & Linn, 1995). This paper presents three general waves of education reform within the Excellence Movement.

The First Wave

Following the input from the first series of reports, the first wave focused on doing "more of the same." The first series of reforms efforts "...did not question the basic structure of education nor the system of which it was a part" (Greenman, 1994, p. 8). This first wave sought excellence through state mandates requiring an intensification of education efforts already in place (Boyd, 1990). Efforts tended to center around raising course and graduation requirements (Boyd, 1990; CPRE, 1989; Firestone, 1990; Lipsky, 1992), more student testing (Boyd, 1990; CPRE, 1989; Feir 1995; Olden & Marsh, 1990), strengthening teacher certification requirements (Firestone, 1990; Lipsky, 1992), and more teacher testing (CPRE, 1989; Firestone, 1990; Olden & Marsh, 1990). During this wave, 45 states had raised or put in place graduation requirements (Firestone, 1990), and 40 states had instituted or increased student testing (Feir, 1995). Teachers did not escape the testing. Career entrance testing was in place in 43 states (Feir, 1995), and teacher testing of some sort for certification was being done in 46 states (CPRE, 1989). These efforts were geared toward increasing the "rigor" in the American education system (Michaels, 1988). Their efforts at reform, however, were soon under attack.

Very top-down and bureaucratic, the reform efforts of the first wave were seen by many educators and scholars as inadequate (Murphy, 1990). Within a few years, reformers felt that no real changes were occurring. Discussions of school culture and change processes started taking place (Greenman, 1994). It was in the mid-1980s that the second wave of commission reports was released spawning the second wave of reform.

The Second Wave

This wave of educational reform efforts is referred to as the restructuring movement (CPRE, 1989; Greenman, 1994). Reform efforts of this wave questioned the very top-down, centralized control, bureaucratic model that had been reinforced through the efforts of the first wave of reforms. Murphy (1990) recalled that these efforts were geared to empower educational professionals and adults and build a "bottom-up" oriented system that made the "people" the real focus of education.

A major area of focus of these reforms was school governance (Boyd, 1990; CPRE, 1989; Murphy, 1990). Emphasis in restructured schools was on "school-site autonomy, shared decisions-making among school staff, enhanced roles from teachers and parents, and regulatory simplicity" (CPRE, 1989, p. 3). A second area of emphasis was in the empowerment of education professionals. Teachers were to be seen and treated as professionals and empowered to make the decisions that were necessary for learning to occur (Boyd, 1990; Murphy, 1990). There was also an interest in reorganizing instruction so that students would engage in more in-depth learning, engaging in higher-order thinking so that students would not only know "how" but also "why" (CPRE, 1989; Michaels, 1989).

Plato (1992) identified five general trends of this second wave of reforms: 1) a movement to outcome-based systems with certification of mastery, 2) a shift toward performance-based assessments, 3) a blurring of the differences in levels of schools, 4) an increase in local control of learning, teaching, and assessment practices, and 5) more "accountability for educational system 'products' resulting from the above choices" (p. 44).

The Third Wave

With the trend toward decentralized control and decision making, questions began being asked about what it was that teachers should teach and students should learn. The crystallization of these thoughts into practice is what is today called standards-based reform (Jennings, 1998; Marzano, Pickering, & McTighe, 1993). The goal was to clearly identify what students should know and be able to do; in other words, establish standards. Manatt (1993) outlines the change in philosophy connected with this effort: 1) set goals or standards, 2) provide the freedom to accomplish the goals, and 3) hold people accountable.

The beginning of standards-based reform is often linked to a historic education summit called by President Bush in 1989 and attended by state governors (Baker & Linn, 1989). The product of that summit was the agreement on a set of national goals for education. Though this garnered much attention, the work already being done by the states on state-level standards most likely set the stage for early acceptance of the national goals by the governors (Pipho, 1996). These goals were endorsed by President Bush in 1990 and enacted in legislation in the America 2000 Act. Strong government involvement continued through the enactment of the Goals 2000: Educate America Act of 1994 signed into law by President Clinton. Goals 2000 called for development of national standards, tests to measure achievement of those standards, and provided aid to those working on or raising standards (Baker & Linn, 1995). The Elementary and

Secondary Education Act reauthorization was also finalized in 1994, and more freedom and flexibility to fashion programs to meet needs was given in exchange for demanding higher standards (Jennings, 1998).

The Three Components of Standards-Based Reform

Standards-based reform is often seen as having three components: standards, assessments, and accountability (Finn & Rebarber, 1992; National Education Association, 1997). These three components "...add a great deal of prescriptiveness, consistency, and power..." (Porter, 1994, p. 444) to the efforts to improve schools.

Standards

Standards are certainly not new to American education. Minimum competency testing programs which swept the nation in the 1970s and 1980s (O'Neil, 1992) required the creation of minimum standards of performance (Baker & Linn, 1995). The institution of the increased standards of the first wave of reforms in the Excellence Movement affected both students and teachers. There are three properties Porter (1994) identified that set the standards of the latest reform efforts apart: 1) they apply to all students, 2) they are criterion-referenced and specify what should be known and done, and 3) they reflect an interest in holding schools accountable for what they produce. A total of 40 states now have standards in all core areas (*Education Week*, 1999) and 49 states have established or are determining common academic standards for students (American Federation of Teachers, 1999).

Assessment

Assessment is also not a new concept in education, but in the standards-based reform efforts, it is critical that assessment be tied to standards (Schmoker, 1996). Baker and Linn (1995) see assessment, in whatever form, as key to bringing meaning to standards and curriculum. It is the "linchpin" to the entire reform effort (Finn, 1995).

Testing has been steadily increasing. Haney & Madaus (1989) estimated that testing has been increasing by between 10 percent and 20 percent a year over the past few decades. Recent increases in testing have been fueled by the waves of education reform. By 1988, all but three states required that local districts test public school students at least once between grades one and 12 (Harnisch & Mabry, 1993). Stake (1999) summarized it, "... formalized student assessment has become the most widely used indicator of school quality" (p. 668).

Because of their efficiency and reliability, multiple-choice, standardized tests have been the dominant method of student assessment (Harnisch & Mabry, 1993). Stiggins (1994) outlines 60 years of standardized testing layering: local scholarship testing in the 1930s, college admissions testing in the 1940s, the distribution of published test batteries in the 1950s, testing for accountability in the 1960s, the surge of statewide testing in the 1970s, the use of testing for national and international comparisons in the 1970s and 1980s, and the current interest in assessing every pupil. Administrators, policy makers, and the public have come to rely on standardized forms of assessment as primary indicators of performance (Moss & Schutz, 1999).

Interest in performance assessment has grown. In part, this is due to concerns over the possible harmful effects of norm-referenced testing (Harnisch & Mabry, 1993; Shepard, 1992). Critics claimed these tests were biased or unfair to some kinds of students (Ayers, 1993; Haney & Madaus, 1989); corrupted the educational processes of teaching and learning (Haney & Madaus, 1989); and were used mostly because of low costs and ease of use (Linn, 1992; Worthen, 1993).

The design of norm-referenced tests was criticized as testing only low-level basic skills (Haney & Madaus, 1989; Wiggins, 1989a) and identifying at least half of students as below average (Ayers, 1993; Wiggins, 1989b). There was also a desire to develop assessment techniques that would allow students to demonstrate their knowledge and skills (O'Neil, 1992).

Accountability

The discussion of accountability in education has been going on for some time, and there is no shortage of literature for those who are interested. Kirst (1990) reported that an estimated 4,000 articles and books had been produced on the topic of accountability from 1969 to 1976. The emergence of the topic in the 1960s was tied to wanting to certify that students could perform certain tasks (Lessinger, 1971) and focused more on outcomes of the education process instead of funding (Tyler, 1971). Stiggins (1994) connected the mastery learning and criterion-referenced testing of the 1960s, the behavioral objectives movement of the 1970s, the minimum competency testing of the early 1980s, and the current reform efforts to a singular accountability movement.

Taking into account the various purposes or models in which accountability could be applied, Kogan (1986) offered the following definition of accountability: "a condition in which individual role holders are liable to review and the application of sanctions if their actions fail to satisfy those with whom they are in an accountability relationship" (p. 25). Kogan also proposed that there were three main types, or models, of accountability: public or state control, professional control, and consumerist control. On a more tangible level, Sciara & Jantz (1972) stated that, "Basically, accountability means that public schools must prove that students at various levels meet some reasonable standard of achievement." Realizing that even after decades of debate accountability means different things to different people, Frymier (1996) offered this short working definition, "To be accountable means to be answerable, to be responsible."

In the 1995–1996 school year, 46 states conducted statewide assessment of students, 38 states reported using statewide assessments for accountability purposes, and 23 states reported administering assessments for student accountability (National Educational Association, 1997). Mitchell (1996) reported that 25 states required school-level profiles. Olson (1999) reported those numbers had risen, with 48 states testing their students and 36 states publishing report cards on individual schools.

There has also been a shift in the focus from inputs to outputs and outcomes (Macpherson, 1995; Manatt, 1993; Porter, 1994). There is also a trend for policy makers to look at these outcomes in order to reward or punish based on the success in efficiently educating children (Olson, 1999). Stake (1999) reported that, "Funding, autonomy, and privilege have been attached..." (p. 669) to scores. Shanker (1995) argued that the standards-based reform efforts would "...be no more successful than any other reform effort unless stakes are attached to the assessments" (p. 152).

The Quest for Quality Data

The use of standards, designing of assessments, and calls for accountability with "high-stakes consequences" created a heightened focus on data. Tests, and other sources of data, become the "...vital link in any goal-based, results-driven accountability system" (Finn, 1995, p. 125). Pipho (1999) described how the standards-based reform movement has started a repeating cycle that "...feeds on and demands better data" (p. 485).

A concern, then, is that in educational assessment, missing data are a widespread problem (Beaton, 1997) and is an "undeniable characteristic" of educational research and evaluation (Velotta, 1995). The problem of missing data is something that affects many social science data sets (Little & Rubin, 1990), tends to have a greater impact in applied settings (Raymond, 1987;

Raymond & Roberts, 1987; Roth, 1994), and can create additional problems in longitudinal studies (Velotta, 1995). Information can be missing for any number of reasons: students absences (Beaton, 1997; Kromrey & Hines, 1991), scheduling problems (Beaton, 1997), transfers of teachers or students (Cohen & Cohen, 1983), lost paperwork (Kromrey & Hines, 1991), or miscoding (Kromrey & Hines, 1991).

The typical solution to the missing data problem is a technique referred to as listwise deletion (Witta, 1994) and is used by many researchers (Roth, 1994). In listwise deletion, all cases with any amount of missing data are eliminated from the data set. It is also the default option for handling missing data on many statistical computer programs (Raymond & Roberts, 1987; Velotta, 1995; Witta, 1994). Even if the percentage of missing data is small for each variable, the number of cases eliminated can become quite large, especially if there are a large number of variables involved (Gleason & Staelin, 1975; Raymond & Roberts, 1987).

The loss of data associated with listwise deletion raises a number of issues for researchers. First, much data may actually get ignored or overlooked (Kromrey & Hines, 1991; Raymond & Roberts, 1987). This reduction in data decreases the power of statistical tests and can introduce bias (Raymond, 1987; Roth, 1994). Listwise deletion, as a method for handling data, can be the worst one to use (Kromrey & Hines, 1991; Raymond & Roberts, 1987). There are many other missing data techniques available to deal with missing data, but those different methods may actually produce different results (Witta, 1994).

So, is there a "best" method that can be used? Sadly, at present, the answer is no. Even though missing data are common, there is little advice on how to handle it (Raymond & Roberts, 1987). Many solutions have been proposed by statisticians, survey researchers, and others, but no one answer has been found (Basilevsky, Sabourin, Hum, & Anderson, 1985; Witta & Kaiser, 1991).

Statement of the Problem

Education finds itself in a reform movement that focuses on standards, assessments, and accountability. In a period where the call for accountability has raised the stakes surrounding student performance, there is a need to investigate the effects missing data techniques can have on results. The Lincoln County School District #1 in Diamondville, Wyoming has devoted several years of effort in curriculum renewal and alignment work. The data in this study come from criterion-referenced assessments in reading, language arts, and mathematics that were developed as part of this work. The problem for this study was to investigate whether the selection of a missing data technique for handling missing data impacts indicators of student achievement.

Pretest, posttest, and gain scores on the tests were used. Tests scores are from the 1995–1996 school year. Student demographic information in the form of gender, socioeconomic status, attendance, and grade were also investigated. An interest in comparisons across grade levels and varying test length by grade led to the creation of a "level of mastery score," or percent indicator, by dividing a score on a particular test by the number of items on the test.

Purpose of the Study

The purpose of this study was to examine variations in the performance of K–12 students on criterion-referenced reading, language arts, and mathematics in relation to the selection of missing data techniques applied to student achievement data. Wilkerson (1997) stated that much of the work done by Lincoln County School District #1 was to help meet new state accreditation standards, especially evidential information called for in the *1995 Wyoming School Accreditation Guide*. This study was of interest primarily for its investigation of how results, and the information gleaned from those results, change as decisions surrounding data treatment, specifically the issue of missing data, are made.

There are two goals of this study. The primary goal was to determine what impact missing data technique selection had on student achievement results. A second goal was to investigate a method to provide assistance to a district in making the selection of a missing data technique.

Objectives of the Study

The following objectives were outlined to accomplish the task of fulfilling the purposes of this study:

1. To investigate the literature on missing data research.
2. To determine if results varied as missing data techniques varied.
3. To investigate a method districts could use to assist in missing data technique selection.

Research Questions

This study addresses two main research questions:

1. Would the use of different missing data techniques affect the results found with the data from Lincoln County School District #1 in Diamondville, Wyoming?
2. How well do different missing data techniques perform?

In order to adequately address the first research question, a series of additional research questions tied specifically to results were also addressed:

a. Did students' level of mastery improve from pretest to posttest?

b. Do gains in level of mastery differ between males and females?

c. Do gains in level of mastery differ between high- and low-SES students?

d. Do gains in level of mastery differ between students with high levels of absence and normal levels of absence?

e. Do gains in level of mastery differ by grade level?

f. Do posttest levels of mastery differ between males and females?

g. Do posttest levels of mastery differ between high- and low-SES students?

h. Do posttest levels of mastery differ between students with high levels of absence and normal levels of absence?

i. Do posttest levels of mastery differ by grade level?

The second research question presented a difficult problem: if data are missing, what criterion can be used to measure performance? As a consequence, this research question was addressed by conducting a simulation using the district's data. The discussion of the results of the simulation will focus on descriptive statistics associated with each missing data technique and deviation indicators where appropriate.

Research Hypotheses

There was a research hypothesis tied to each of the result-oriented questions asked to address the first research question:

a. Level of mastery scores will be significantly higher than pretest level of mastery scores in all subjects.

b. There will be no difference in gain in level of mastery between males and females in any subject.

c. There will be no difference in level of mastery between high- and low-SES students in any subject.

d. There will be no difference in gain in mastery between students with high levels of absence and normal levels of absence in any subject.

e. There will be no difference in gain in level of mastery between grade levels in any subject.

f. There will be no difference in posttest levels of mastery between males and females in any subject.

g. There will be no difference in posttest levels of mastery between high- and low-SES students in any subject.

h. There will be no difference in posttest levels of mastery between students with high levels of absence and normal levels of absence in any subject.

i. There will be no difference in posttest levels of mastery between grade levels in any subject.

Basic Assumptions

1. Increased interest in standards, assessments, and accountability leads to a focus on data.
2. The quality of information from which conclusions are drawn is tied to how data are treated.
3. Districts are interested in treating all stakeholders in the education process as fairly as possible.
4. All districts do not possess the statistical expertise to apply cutting-edge procedures developed in statistics and survey research.

Delimitations of the Study

1. The data for this study were from Lincoln County School District #1 in Diamondville, Wyoming during the 1995–1996 school year.
2. It is assumed that the criterion-referenced assessment system has been deemed adequate for assessing student achievement and that the administration of these tests was conducted appropriately.

3. Only data from Lincoln County School District #1 was used and the associated patterns and relationships in the data may limit the generalizability of the results.
4. Only five missing data techniques were reviewed in this study.

Human Subjects Release

The data in this study were used with the full knowledge and permission of Lincoln County School District #1. Iowa State University's Committee on the Use of Human Subjects in Research also reviewed this study and concluded that any risks were outweighed by the potential benefits and that the welfare of the human subjects was adequately protected.

CHAPTER II. REVIEW OF LITERATURE

Information for this study primarily came from two areas: education and survey research. The context for the study lies in the call for more educational accountability and higher demands on data. The tools to evaluate the performance of the missing data techniques in this study were taken from work done in the field of survey research which has a body of knowledge on missing data techniques. The search for information started with keyword searches in the *Educational Resources Information Center*, Iowa State's Parks Library, and the Internet. Books, papers presented at conferences, dissertations, articles, and contact with faculty at the Statistics Department at Iowa State University allowed for a narrowing of focus to useful information.

The Need

The components of standards-based reform, standards, assessments, and accountability, are the same three components of what has been called the "new educational accountability" (Elmore, Abelman, & Fuhrman, 1996). In order to make a "complete" accountability system, Newman, King, and Rigdon (1997) add a fourth component: a body that reviews data, judges against standards, and hands out sanctions and rewards. These systems are increasingly focusing on educational outcome measures, with a primary emphasis on student performance (Elmore, Abelman, & Fuhrman, 1996; Meyer, 1997). With demands on student performance data increasing, it is not unreasonable to invest some time addressing the quality of the data.

An area of concern in regard to the quality of data in statistical analyses, especially when results are tied to major decisions or policy, is that of missing data (Henry, 1995). Social and behavioral science studies "frequently suffer" from missing data (Little & Schenker, 1995). The existence of incomplete data in educational research and evaluation is an "undeniable characteristic" of the work (Velotta, 1995) and is "widespread" in educational assessment (Beaton, 1997).

Meyer (1997) described average test scores aggregated at various levels as "contaminated" due to student mobility, which becomes a greater problem as one moves from national results down to the building level. In discussing accountability in South Carolina, Clotfelter and Ladd (1996) reported that the size of the database used to track students from fourth grade to fifth grade fell from a total of 47,000 students to a usable set of matching results for 41,650, a drop of approximately 11.4 percent in one year's time. A total of 2.4 percent of the data was dropped due to missing or incomplete data and the remaining 9 percent were dropped because students had changed schools. A recent example of a debate surrounding missing data outside of education is how the year 2000 census should be conducted.

The 1990 census is believed to have undercounted the population of the United States by approximately four million people, or around 1.6 percent (Leung, 1998; Maxwell, 1997). The Census Bureau outlined procedures to conduct sampling instead of enumeration in order to provide a more accurate count for the 2000 census. This became a political issue because those missed by traditional census methods, minorities in particular (Leung, 1998), were identified as more likely to vote Democratic (Williamson, 1998). With Republicans arguing for strict enumeration and Democrats arguing for sampling, a series of court battles followed with the final decision left to the United States Supreme Court (Biskupic, 1998). In a divided ruling, the United States Supreme Court ruled in January of 1999 that the 2000 census could not be adjusted through the use of sampling (Associated Press, 1999).

Potential Problems

Little and Schenker (1995) outline three major problems that can be created by missing data: 1) analyses that ignore missing data characteristics may introduce bias to the results, 2) statistical estimates may be less accurate due to information loss, and 3) statistical procedures designed for use on complete data sets are affected making analyses more complicated. With the current focus on outcomes, there is another important issue: the condition of the results.

With missing data so prevalent, decisions concerning its treatment occur all the time. Whether conducting research or computing analyses, decisions must be made to deal with missing data. However, this decision is often an unconscious one (Velotta, 1995) being left to statistical software default settings. At issue, however, is that different missing data techniques can lead to differing results (Velotta, 1995; Witta & Kaiser, 1991). In a review of three different studies investigating the High School and Beyond data set, Ward and Clark (1991) found that the choice of missing data technique used did impact results. The routine nature with which decisions about missing data are made and the complexity of issues generated by the appearance of missing data led Patton (1982) to state in his book on evaluation, "The decision on how to handle missing data should be a conceptual one rather than a routine one" (p. 256).

Definitions and Concepts

The following definitions will be used in this study:

Cell-mean substitution: This missing data technique (MDT) involves the segmenting of the data set into subsets, or cells, and the replacement of missing values with the variable mean that was computed using the complete cases within that cell. It is sometimes referred to as conditional-mean substitution. It is one of the MDTs examined in this study.

Cold-deck imputation: Similar to hot-deck imputation, this MDT substitutes values taken from outside the data set for missing values within the data. This MDT is not examined in this study.

Complete case: When a subject in a data set has values for all variables in the set.

Grand-mean substitution: This MDT involves the replacement of missing values with the variable mean that was computed using the available cases. This technique is also referred to as unconditional mean substitution. It is one of the MDTs examined in this study.

Hot-deck substitution: This MDT involves the matching of an incomplete data record with a complete record having similar characteristics and substituting the existing value for the missing value. Like cell-mean substitution, this MDT usually involves the segmenting of the data set into subsets. The selection of a value from existing values within a cell can be performed in a number of ways. This MDT is not examined in this study.

Incomplete case: When a subject in a data set fails to have values for all variables in the set.

Iterative regression: This MDT is much like the multiple regression technique. Upon completion of the first iteration, however, new regression equations are generated and new estimates for the missing values are produced. This process continues until there is little to no difference in the estimates. This MDT is not examined in this study.

Listwise deletion: This MDT involves removing all cases for which there are any missing data. It is also referred to as the complete-case analysis technique. It is one of the MDTs examined in this study.

Mean absolute deviation score: This deviation measure is calculated by summing the absolute value of the differences between all actual and estimated, or imputed, values and dividing by the number of deviations. Because this value is always nonnegative, it is used to measure how "close" imputed values are to actual values.

Mean deviation: This deviation measure is calculated by summing the differences between all actual and estimated, or imputed, values and dividing by the number of differences. It is used as an indicator of bias in imputed scores.

Missing data technique (MDT): Any process followed to address the issue of incomplete or missing data. The abbreviation MDT will be used throughout this work.

Multiple imputation: This MDT involves the generation of more than a single value for a missing datum. The data are used to create multiple data sets from which statistics and conclusions are drawn. This MDT is not examined in this study.

Multiple regression: This MDT involves the use of listwise deletion to create the initial correlation matrix. Then each missing value for a variable is replaced by an estimate determined by regressing the missing variable on the non-missing variables. This MDT is not examined in this study.

Pairwise deletion: This MDT uses all available data in the computation of means and variances, and all available pairs of values for the computation of correlations. This technique is also referred to as the available-case analysis technique. It is one of the MDTs examined in this study.

Power: The ability of a statistical test to detect a relationship.

Root mean square deviation score: This deviation measure is calculated by finding the square root of the average of the squared differences between all actual and estimated, or imputed, values. Because this value is always nonnegative, it is used to measure how "close" imputed values are to actual values.

Simple regression: This MDT involves the use of listwise deletion to create the initial correlation matrix. Then each missing value for a variable is replaced by an estimate determined by regressing the missing variable on the one variable with which it has the highest correlation.

If that variable is also missing, the procedure continues on to the next highest correlated variable. It is one of the MDTs examined in this study.

Stochastic regression: A variation of simple or multiple regression in which an error term is added to the estimate. This error term is often in the form of a random addend.

Zero-order techniques: A term used to refer to techniques that do not utilize correlational information in dealing with missing data. This term often refers to listwise deletion, pairwise deletion, and grand-mean substitution.

General Approaches to Missing Data

Taxonomies of missing data techniques are varied. Roth (1994) splits the techniques into two groups: 1) "simple," in which listwise deletion, pairwise deletion, mean substitution, regression and hot-deck procedures are placed, and 2) statistical model methods in which methods requiring knowledge of the underlying distributions are needed. Little and Schenker (1995) also divide missing data techniques into two groups. The "naive" approaches include complete-case analysis (listwise deletion), available-case analysis (pairwise deletion), and unconditional mean (grand mean) imputation. Their more "principled" approaches included conditional mean (cell mean) imputation, regression, and hot-deck techniques as well as multiple imputation and model based methods. Little and Rubin (1990) outlined three main approaches to handling missing data: imputation which included all techniques designed to replace a missing value, weighting which included complete-case analysis and other procedures for handling unit nonresponse, and direct analysis of the incomplete data which included available-case analysis and model-based approaches.

Thinking in terms of what is done with the data, Kromrey and Hines (1991) outlined two fundamental approaches: 1) where missing data are not included, which included listwise and pairwise deletion, and 2) where missing data are estimated, which included all techniques designed to create a complete data set. Beaton (1997) matched this last breakdown, though labeled the two groups as "commonly available" and "imputation." For the purposes of this chapter, two major categories of missing data techniques will be discussed: deletion techniques and common imputation techniques.

Techniques Investigated in This Study

As mentioned in the definitions of various missing data techniques, this study focuses on five techniques. Listwise deletion was included because of its status as the most commonly used technique (Raymond & Roberts, 1987) and the default in many statistical programs (Little & Schenker, 1995). Pairwise was included in the study because it could be applied to nonrectangular data sets (Little & Rubin, 1987). Three imputation techniques were included in the study: grand-mean substitution, cell-mean substitution, and simple regression imputation.

Grand-mean substitution was selected because of its status as one of the most commonly used imputation techniques (Hegamin-Younger & Forsyth, 1998; Velotta, 1995). Cell-mean substitution was included because of its supposed improvements over grand-mean substitution (Little & Schenker, 1995). Because research showed that the performance of different regression techniques varied little (Raymond & Roberts, 1987), simple regression was selected for ease of use. This selection of missing data techniques matches recommendations in the literature in terms of the number and makeup of techniques to use (Basilevsky, Sabourin, Hun, & Anderson, 1985; Raymond & Roberts, 1987; Velotta, 1995; Ward & Clark, 1991).

The discussion of missing data techniques that follows pertains to techniques in this study as well as others mentioned in the literature.

Deletion Techniques

Listwise deletion

Listwise deletion, sometimes referred to as complete-case analysis, is a missing data technique that omits all cases with partial data. Listwise is the most obvious approach to handling missing data (Raymond & Roberts, 1987) and is used by most researchers (Roth, 1994). Listwise deletion is the default option on most statistical software packages (Little & Rubin, 1990; Little & Schenker, 1995; Raymond, 1987; Raymond & Roberts, 1987; Roth, 1994; Witta & Kaiser, 1991). This is in line with the fact that standard statistical procedures have been developed to work with complete rectangular data sets (Little & Rubin, 1987). For listwise deletion to perform well, it should be used when the cases that were eliminated were similar to those with complete data in regard to the variables of interest (Beaton, 1997; Little & Schenker, 1995; Malhotra, 1987; Witta & Kaiser, 1991).

The advantages of listwise deletion are its ease of use (Little & Rubin, 1987, 1990; Little & Schenker, 1995); the comparability of statistics because of the use of a single common set of data (Little & Rubin, 1987); and the ability to use standard statistical procedures without additional work (Little & Rubin, 1987). Disadvantages include the sacrifice of useful data (Little & Rubin, 1990; Raymond, 1987; Raymond & Roberts, 1987; Roth, 1994; Witta & Kaiser, 1991), decreased power (Raymond, 1987; Roth, 1994; Roth, Campion, & Jones, 1996), and possible bias, especially if deleted cases differ from remaining cases (Beaton, 1997; Little & Rubin, 1990; Little & Schenker, 1995; Raymond, 1987; Roth, 1994).

Pairwise deletion

Pairwise deletion, sometimes referred to as available-case analysis, is a technique that uses all available information in the calculation of statistical results. All values of a variable are used in the calculation of descriptive statistics like mean and variance, and all pairs of available values are used in the calculation of correlations. A popular alternative to listwise deletion (Little & Rubin, 1987; Raymond & Roberts, 1987), it is based on the assumption that using the maximum amount of available data will produce better statistics (Witta & Kaiser, 1991).

The advantages of pairwise deletion advantages are that it can be applied to nonrectangular data sets (Little & Rubin, 1987), and it preserves a great deal of data (Beaton, 1997; Little & Schenker, 1995; Roth, 1994; Roth, Campion, & Jones, 1996). Disadvantages tied to the use of pairwise deletion are changing samples used in calculating and applying statistics (Little & Rubin, 1987; Raymond, 1987; Velotta, 1995) and inconsistencies in statistics dealing with the relationship between variables such as correlations and regression weights (Beaton, 1997; Little & Schenker, 1987; Malhotra, 1997; Raymond, 1987).

Imputation Techniques

As a general approach, imputation involves the "filling in" of missing data as opposed to ignoring data as in the deletion methods. This "filling in" is done with the hope of producing unbiased statistical estimates, maintaining the natural shape of distributions, and consistency in the relationship between different variables (Armoogum & Madre, 1998). The advantage to using imputation techniques is that a rectangular set of data is created to which standard statistical analyses can be performed (Beaton, 1997; Little & Rubin, 1987, 1990; Little & Schenker, 1995).

Grand-mean substitution

Grand-mean substitution is among the most common of imputation techniques (Hegamin-Younger & Forsyth, 1998; Little & Rubin, 1990; Little & Schenker, 1995; Raymond & Roberts, 1987; Velotta, 1995). It is also referred to as the unconditional mean substitution. Grand-mean substitution involves the substitution of any missing value with the overall mean for that variable. Its advantage is that it is simple to apply (Hegamin-Younger & Forsyth, 1998; Little & Rubin, 1987, 1990; Roth, 1994). The disadvantages of using grand-mean substitution are that it overestimates sample size (Velotta, 1995), underestimates variance (Little & Rubin, 1987, 1990; Raymond, 1987; Roth, 1994; Roth, Campion & Jones, 1996; Velotta, 1995; Witta & Kaiser, 1991), underestimates correlations (Little & Schenker, 1995; Roth, Campion & Jones, 1996; Velotta, 1995; Witta & Kaiser, 1991), distorts the shape of distributions (Beaton, 1997; Little & Rubin, 1987; Velotta, 1995), and can impact statistical tests of inference (Armoogum & Madre, 1998; Little & Schenker; 1995). An example of its application would be to use the average math pretest level of mastery, or percent correct, across all grade levels for any student with a missing math pretest score.

Cell-mean substitution

A technique seen as an improvement over grand-mean is cell-mean substitution (Little & Rubin, 1990; Little & Schenker, 1995). This technique is also called conditional mean substitution. Cell-mean substitution breaks the data sets into "cells," usually based on categorical variables within the data set. Within-cell means are calculated and substituted for missing values within the cell. This technique is particularly useful when one of the variables of interest is categorical (Little & Rubin, 1990; Velotta, 1995). The disadvantages of cell-mean substitution are the same as grand-mean substitution, except that the effects are lessened (Little & Rubin, 1987, 1990; Velotta, 1995). In terms of the current study, grade level was used to divide the district's data into 13 cells, one for each grade kindergarten through 12th grade.

Hot- and cold-deck imputation

Using the creation of cells to help determine values to be imputed, hot- and cold-deck imputation are techniques that rely on values provided under matching circumstances. After the data set has been divided into cells, hot-deck imputation replaces missing values within a cell with a value of the variable that already exists within the cell. The existing value within the cell can be chosen in a number of ways: randomly, sequentially, or by any number of functions. Cold-deck imputation operates similarly to hot-deck imputation except that it pulls its values from a source other than the current data set. Proponents claim these techniques are better because realistic data are used (Little & Rubin, 1987; Roth, 1994) and provide better estimates of variation because more than one value is imputed (Little & Rubin, 1990). Difficulties include the number of categories can become unmanageable and the categorizing of variables that can sacrifice information (Roth, 1994). More empirical work is needed to determine the accuracy of the techniques (Roth, 1994).

Regression techniques

Regression techniques are a collection of methods that are being tried by an increasing number of researchers (Roth, 1994). In general, missing values are substituted with an estimate that is generated from an equation that regresses the missing variable onto one or more covariates. The techniques can vary in a number of ways: the starting treatment of the data (use of listwise, pairwise, etc.), the number of covariates used in the equations, whether an iterative approach is used, or if a stochastic element is added to the equation (Raymond, 1987). Work comparing the various techniques has tended to show little difference in performance (Raymond & Roberts, 1987).

The use of a regression technique tends to save more data than pairwise deletion, preserve the variation and shape of a distribution, and lessen correlation to a lesser degree than grand-mean substitution (Roth, 1994). Studies have tended to start the process by calculating an initial regression equation from complete cases (Little & Schenker, 1995; Roth, 1994) and then use that equation to generate a substitute for missing values. Researchers should be aware that the linear relationships built into regression may intensify the factor structure among the variables (Roth, 1994) and increase correlations among filled in values (Beaton, 1997). Slight underestimation of variances may occur based on the number of similar groupings of values fed into the regression equations (Beaton, 1997; Roth, 1994).

An option that can be added to the use of regression equations is that of an error term. Confronted with the possibility of imputing the same value a number of times, an error term that varies in value can be added to the equation. This is sometimes called stochastic regression. The goal is to counteract the attenuation of the variance (Beaton, 1997) with the values of the error term often having a mean of zero and a variance which matches the circumstance (Little & Rubin, 1987).

Concern over a single application of a missing data technique has been handled through increasing the number of replications and through the use of an iterative regression process. In iterative regression, an initial regression equation is generated and used to impute missing data. With the new complete set of data, a new regression equation is generated. This process is continued until the regression weights show very little change (Roth, 1994).

Comparison Studies Using Computer Generated Data

Haitovsky (1968) investigated the impact of listwise deletion and pairwise deletion on regression analysis. Eight data sets with preset means, variances, and correlations were generated and computer subroutines used varying patterns to create missing observations. A total of 10 runs were performed for each deletion pattern that was used. Results from the data sets treated with listwise deletion and pairwise deletion were compared to the known regression equations generated from the complete set. The listwise deletion method was found to be superior to pairwise deletion in the estimation of regression weights. It was also recommended that more than one method of handling missing data should be applied if the proportion of missing data was high.

Timm (1970) looked at the impact several missing data techniques had on variance-covariance and correlation matrices. The number of techniques investigated increased to four, with listwise deletion, pairwise deletion, grand-mean substitution, and multiple regression being used. Matrices with known variance-covariance and correlation structures were generated and used for comparison. Incomplete data matrices were created with one percent, 10 percent, and 20 percent of the data randomly deleted. The effects of varying sample size, number of variables, percent missing data and average intercorrelation of variables were examined. It was found that none of the four missing data techniques performed uniformly best in estimating either the variance-covariance matrix or the correlation matrix. It was noted that the grand mean and listwise deletion methods did not perform as well as the other techniques in almost all cases that were studied and that the multiple regression method performed less well as the intercorrelation of variables fell.

A study which revisited the comparison of listwise and pairwise deletion methods was done by Kim and Curry (1977). The data consisted of a computer generated simulation of 1,000 samples from which a set correlation matrix was created and used as the population model. A random deletion of 10 percent of the cases for each variable was then performed. This sampling was repeated 10 times with the resulting sample correlation and covariance matrices compared with the population model. The results found that pairwise deletion performed better than listwise deletion because it had smaller deviations from the model.

Basilevsky, Sabourin, Hun, and Anderson (1985) conducted an involved study of the impact of several missing data techniques in linear regression models. Among the techniques investigated were listwise deletion, pairwise deletion, grand-mean substitution, multiple regression, and iterative regression. The experiment consisted of drawing samples from a number of pre-defined multivariate normal populations, each defined by a combination of factors. The factors of multicollinearity (three levels), explained variation (three levels), and sample size (two sizes) created 18 conditions that were each replicated 10 times resulting in 180 sets of data. Each of these data sets was then passed through three different missing data filters, removing 10 percent, 30 percent, and 50 percent of the observations making a total of 540 sets from which regression estimates were generated. The criteria consisted of the explained variance and a mean square error for the beta weights. The results varied by the amount of data that was missing. For 10 percent missing data most of the estimators performed well. For 30 percent missing data listwise deletion, pairwise deletion, and grand-mean substitution were best for large and small sample size. For 50 percent missing data the multiple regression and iterative multiple regression performed best for small sample size and listwise deletion, pairwise deletion and mean substitution performed better for large sample sizes. The researchers recommended that use of several of the first order estimators (listwise, pairwise, and grand mean) and to advance to other techniques if the first order techniques offer varying results or if large amounts of data would be lost as a result.

Raymond and Roberts (1987) investigated the impact of listwise deletion, grand-mean substitution, simple regression, and iterative regression on regression equations used in selection research. Samples of multivariate data were generated by computer to deliver correlation matrices with predetermined values. These samples were used for purposes of comparison. Four different variables were used to generate the data: correlation matrix (four options), sample size (three options: n = 50, 100, 200), percent of missing values (three options: 2 percent, 6 percent, 10 percent), and missing data technique (four options). For each correlation matrix there were 36 different conditions, each of which was replicated 30 times. The criteria used were closeness of the amount of variation explained and a mean square error of the beta weights.

Differences in mean square errors for the beta weights tended to increase as sample size decreased and the amount of missing data increased with listwise deletion being the least accurate followed by the grand-mean substitution technique. Differences between the two regression techniques were consistent, yet usually small. All of the missing data techniques in the study produced close estimates of the explained variation. Overall, the two regression techniques provided the most accurate regression equations, with listwise deletion being the least accurate. The researchers recommended that if more than 5 percent of values were missing from any single variable, listwise deletion (the default in most cases) should be compared to at least one other missing data technique. Should differences among the techniques occur, correlational methods should probably be used. Given the close performance of the two regression techniques, even though the iterative regression required considerably more work, the researchers suggested the simple regression technique may have more utility.

In a study focusing primarily on regression techniques, Kaiser and Tracy (1988) investigated how well means, predicted values, and covariance matrices were estimated. A series of data matrices were generated by computer that would generate a specific correlation matrix for comparison. Three factors were then used to create the data sets for the study: sample size (three options: 30, 60, 120), the percent of incomplete records (three options: 10 percent, 20 percent, 30 percent), and the number of missing values in an incomplete record (four options: one, two, three, four). Each of the 36 cells was replicated 500 times. Suitability criteria in this study involved discrepancy from means, root mean square error between actual and predicted values, and root mean deviation of covariance elements.

Kaiser and Tracy found that all of the techniques produced unbiased estimates of means. Though not significant enough to rank, grand-mean substitution produced worse estimates of means than the regression variations. The discrepancy between estimated and actual means increased as the amount of missing data increased, though this was decreased as sample size increased. Regression variations were also found to be closer to the original covariance structure and produced better estimates of missing values. The researchers recommended, however, that grand-mean substitution should be used when the purpose is the estimation of means because of its ease of use.

Discussion of Comparison Studies Using Computer Generated Data

In each of the studies using computer generated data, a set of data was used as the population, or "true," set of values. This facilitated the creation of criteria that helped researchers determine the "most accurate" missing data technique. The criteria usually were in the form of a deviation indicator which highlighted the closeness of the missing data technique results with that of the population results. The underlying belief through all of these studies was that the "best" technique was the one that got the "closest" to whatever was being examined. In order to avoid results connected to how data were removed, all of the studies replicated the conditions. The number of replications ranged from 2 to 500, though only one study used more than 30, with 10 being the most common choice. For the most part, the studies investigated the impact of missing data techniques on some aspect of regression.

There were conflicting results on the use of listwise and pairwise deletion. The results of Haitovsky (1968) are in opposition to the results of Kim and Curry (1977). Haitovsky found listwise deletion superior to pairwise deletion, while Kim and Curry found pairwise deletion performed better than listwise deletion. It should be noted that the studies used different criteria: Haitovsky examined regression weights while Kim and Curry examined correlation and covariance matrices. Listwise deletion, often the default approach, has been shown to be less accurate than other forms (Timm, 1970; Raymond & Roberts, 1987). And yet, zero-order techniques have been shown to produce better estimates under some conditions (Basilevsky, Sabourin, Hun, & Anderson, 1985).

All missing data techniques appear to produce good estimates of means (Kaiser & Tracey, 1988), yet have varying impacts on items such as explained variance, regression estimates, and beta weights. In general, there is no one missing data technique that appears to rise above the rest, and a comparison of the results from more than one of the simpler techniques (listwise deletion, pairwise deletion, grand-mean substitution) should be done first. Should there be major differences in these results, more complicated regression procedures may be necessary (Basilevsky, Sabourin, Hun, & Anderson, 1985; Raymond & Roberts,1987).

Table 1 presents a summary of the comparison studies that used computer generated data. The table lists who and when the study was conducted, the missing data techniques investigated, and a brief listing of major results.

Comparison Studies Using Actual Data

Using data collected in 1978 by the Income Survey Development Program Research Panel , Kalton (1983) investigated the impact of several imputation techniques. Starting with cases that had data for all variables of interest (listwise deletion), missing values were created in the variable

Table 1. Summary of comparison studies using computer generated data

Author(s)	Methods compared	Results
Haitovsky (1968)	Listwise deletion Pairwise deletion	Listwise superior in estimating regression weights.
Timm (1970)	Listwise deletion Pairwise deletion Grand-mean substitution Multiple regression	No uniformly best technique, listwise and grand mean techniques perform less well than others.
Kim & Curry (1977)	Listwise deletion Pairwise deletion	Pairwise deletion performs better than listwise deletion.
Basilevsky, Sabourin, Hun, & Anderson (1985)	Listwise deletion Pairwise deletion Grand-mean substitution Multiple regression Iterative regression Other regressions	All satisfactory with 10% missing; Listwise, pairwise, and grand-mean best with 30% missing; and with 50% missing regression based is best for small samples and listwise, pairwise, and grand mean are best for best for large samples.
Raymond & Roberts (1987)	Listwise deletion Grand-mean substitution Simple regression Iterative regression	Regression based techniques provide most accurate regression equations followed by grand mean. Listwise was least accurate.
Kaiser & Tracey (1988)	Grand-mean substitution Simple regression Two-variable regression Other regressions	All produced unbiased estimates of means. Regression based techniques perform better in replicating data structures.

tracking hourly rate of pay by randomly removing to match patterns of missing data in the full sample.

A total of eight different imputation techniques were used: grand mean substitution, cell-mean substitution using eight cells, cell-mean substitution using 10 cells, random hot-deck imputation using the same eight cells previously outlined, random hot-deck imputation using the same 10 cells previously outlined, multiple regression, and two models of multiple regression with error adjustments. In order to avoid issues connected to a single random removal of data, 10 simulated data sets were created and results were averaged across the 10 sets. Those techniques involving randomness in their process were additionally averaged over 10 iterations on each simulated data set.

Using the beginning set as a comparison set, performance was measured through the use of three deviation measures. The mean deviation was used to measure the bias in the techniques, that is to say, whether the techniques over or under estimated the actual values. Mean absolute deviation and root mean square deviation measures were used to indicate how close the techniques reconstructed the data.

The study revealed that all of the techniques had negative mean deviations and underestimated the actual values. The grand-mean substitution technique had the greatest underestimation. On average, cell-mean techniques had the smallest underestimation, performing better then the hot-deck random and multiple regression procedures.

In terms of "closeness" to the actual values, investigation of the mean absolute deviation and the root mean square deviation showed three trends. The techniques which predicted a value without the use of random terms were best (both cell-mean techniques and the multiple regression with the error term). Overall, the worst performing technique was grand-mean substitution. Lastly, those techniques which used 10 cells performed better than those that used eight cells. Kalton also noted that the use of grand-mean substitution and cell-mean substitution techniques created spikes in the distribution of scores at the imputation cell mean values.

In a study that used data from a published article investing success in foreign language training, Raymond (1987) studied the impact of listwise deletion, pairwise deletion, and an iterative regression technique. Of particular interest was how the regression coefficients and amount of explained variation differed with the use of the three missing data techniques. There was no set available for comparison, and each technique was applied once.

The size of the data set resulting from the application of listwise deletion fell from the original 279 cases to 174, leaving 62 percent of the original set. The pairwise deletion samples ranged in size from 197 to 278, with approximately 9 percent missing overall. For the regression procedure, listwise deletion was used to reduce the data set into a usable form leaving 230 cases. Even with this, close to 6 percent of the 230 cases still had missing values. The regression technique was used to impute values in these 230 cases.

After the missing data techniques were applied, regression equations were generated. The equations were used to identify significant predictors of college foreign language grades. The equations generated from the data treated by listwise deletion and regression identified the same five significant predictors with only minor differences in the regression coefficients. There was, however, a notable difference in the amount of variation explained by the two techniques. The data treated with the pairwise deletion technique generated a regression equation that explained a similar amount of variation as the equation generated using data treated by the regression technique, but only identified four significant predictors.

In discussing recommendations, it was noted that no missing data technique was a reliable replacement for a complete data set. Further, Raymond warned that even though listwise and pairwise deletion techniques were sometimes appropriate, they should not be relied on simply because they are easy to use.

In their investigation of listwise deletion, grand-mean substitution, simple regression, and iterative regression, Ward and Clark (1991) reviewed three published articles using data from the 1980 High School and Beyond study. The three studies investigated the impact of public versus parochial schools on students, each using different variables. Of the approximately 28,000 seniors in the database used in the studies, the amount of data used ranged from 86 percent in one study when grand-mean substitution was used for missing independent data to 46 percent in the study applying listwise deletion. The third study made no explicit mention of a missing data technique and used 64 percent of the available cases.

Starting with the same 28,000 students, Ward and Clark matched the three sets of analyses used in the previous studies with the four missing data techniques. With each analysis applied to data treated with four different missing data techniques, this resulted in a total of 12 analyses that were performed without replication. This meant that the original missing data technique was applied in each analysis as well as three additional techniques.

For the study originally using listwise deletion, all three additional missing data techniques produced a larger impact for private schooling. The study utilizing the grand-mean substitution saw the original conclusion of no difference change to support a positive effect for private schools. In the third study, it was found that the effect of private schooling may have been stronger than originally determined.

Ward and Clark discussed a number of conclusions and recommendations. First, that the selection of a missing data technique can impact the statistics as well as the conclusions of a study. In addition, researchers using the default method (listwise deletion) may not fully understand the effect the missing data can have. Reviewing the statistics from across the three studies, the researchers expressed concern over the grand-mean substitution technique's tendency to reduce correlations and suggested that it not be used to estimate missing data. In regard to regression procedures, the close performance of the simple and iterative techniques led to the recommended use of simple regression because of its easier application. In terms of an overall procedure, Ward and Clark recommended that listwise deletion and simple regression both be used and the results compared. Similarities would lead to confidence in conclusions and differences would call for further investigation.

Witta and Kaiser (1991) examined four missing data techniques: listwise deletion, pairwise deletion, grand-mean substitution, and multiple regression. Data were taken from the 1987 General Social Survey conducted by the National Opinion Research Center. After listwise deletion was used to create a sample with no missing values for the criterion variable, a sample of 829 was randomly divided into samples of 414 and 415. Again using listwise deletion, the sample of 414 was reduced to 283 and set aside to be used for comparison. It was assumed that if values were missing randomly, this smaller set represented a random sample of the larger one. The sample of 415 was treated by the four missing data techniques and a regression equation was produced from each set to predict the criterion variable.

A total of five random samples of size 25 and five random samples of 50 were taken from the sample of 414. These samples were used as the "comparison" set. Each of the four regression equations were used to predict the criterion variable in each of the 10 samples.

The only technique different from the criterion in the samples was the grand-mean substitution technique. Witta and Kaiser concluded that the selection of listwise deletion, pairwise deletion, and regression techniques did not affect the effectiveness of regression equations to predict a criterion variable. However, with a note that the differences were present even in small samples, researchers concluded that grand-mean substitution was the most inappropriate way to handle missing data if the goal was to predict the value of a criterion variable.

In a study which used three large sets of field data, Kromrey and Hines (1991) investigated the performance of five missing data techniques. The researchers compared the impact of listwise deletion, pairwise deletion, grand-mean substitution, simple regression, and multiple regression on the regression coefficients and explained variation of a two predictor regression equation.

From each of the three data sets, 100 samples of size 50, 100, and 200 were taken with replacement. For each sample, a proportion of scores were randomly removed in one of the predictor variables. A total of 100 samples were investigated, at six different levels of missing data, from 10 percent up to 60 percent in 10 percent increments. Finally, 100 samples were examined with no missing data and used for comparison. The experiment, then, was a three (data sets) by three (sample size) by six (missing data levels) by five (missing data techniques) design. Results were examined in terms of explained variation, the coefficient for the predictor variable with missing data, and the coefficient without missing data.

In terms of effect on explained variation, differences in effectiveness were evident, differences among the missing data techniques increased as the proportion of missing data increased, and the effects of missing data and their treatment appeared to be stable across sample size. The coefficient for the predictor variable with missing data was consistently underestimated by grand-mean substitution and overestimated by multiple regression. The remaining missing data techniques showed no consistent results. For the coefficient of the predictor variable without missing data, there was a reversal, with multiple regression underestimating and grand-mean substitution overestimating consistently.

Overall, the two deletion methods provided better estimates of explained variation and both regression coefficients. The three imputation methods did not perform well, even when as little as 10 percent of the data were missing. Kromrey and Hines recommended that a criterion should be used in order to examine effects on the parameters being interpreted when evaluating the effectiveness of missing data techniques, even though no tests for significance exist for the comparison of matrices. In addition, it was recommended that evaluation of missing data techniques should occur within the context of realistic data and situations.

In an investigation of six different missing data techniques, Thran and Gillis (1992) used data from the 1991 spring survey of the American Medical Association's Socioeconomic Monitoring System. The survey had traditionally suffered from item nonresponse to questions concerning annual net income and total practice expenses with item nonresponse rates at approximately 23 percent and 35 percent respectively. The researchers investigated the impact listwise deletion, grand-mean substitution, random hot-deck imputation, shortest distance hot-deck imputation, multiple regression, and multiple regression with a random error term in a two-part study.

The first part of the study looked at the mean, median, standard deviation, and percentile scores resulting after using each of the six missing data techniques. Both the hot-deck techniques and the cell-mean technique used 20 cells for income and 30 cells for expenses. Each missing data technique was applied once to the data set and statistics were generated. With the exception of the grand-mean technique, the imputation techniques produced slightly higher estimates of mean net income and expenses than listwise deletion. In terms of the distribution of income percentiles, the hot-deck techniques and the regression with error term technique produced distributions close to that produced using listwise deletion.

The second part of the study was a simulation where a full set of responses was used to generate the same statistics as in the previous study. In addition, regression equations were generated to predict physician earning. A process was used to remove scores in a fashion that resulted in a set that matched nonresponse rates in the full survey. A total of 250 mock samples were created. A set of statistics and a regression equation were generated for the complete set for comparison purposes.

The listwise deletion and grand-mean techniques underestimated the actual means for income and expenses while the reaming techniques overestimated them. With the exception of the regression with error term technique, all techniques produced values with a 95 percent confidence interval of the actual means. Listwise deletion produced lower estimates of median income while the remaining techniques overestimated the median income. With respect to the regression coefficients, the listwise deletion technique produced means that were very close to the actual sample. The grand-mean substitution method was found to produce smaller coefficients in absolute value than the actual sample. Explained variation was lower than in the actual sample for all but the multiple regression technique, and particularly low for the grand-mean and hot-deck techniques.

Overall, Thran and Gillis could not identify a best technique. The hot-deck and regression with error term techniques produced reasonable sample statistics but were not as good as listwise in maintaining the relationship among variables.

Velotta (1995) investigated the performance of four different missing data techniques using kindergarten student performance in language skills. The data were 1993–1994 pre- and posttest scores on the Peabody Picture Vocabulary Test for 2,697 kindergarten students. After selecting variables of interest, listwise deletion was used to create a set of 443 cases. Listwise deletion, grand-mean substitution, cell-mean substitution, and simple regression were applied in two different studies.

The first study started with the set of 443 complete cases and calculated means, standard deviations, and a two by two analysis of variance to test for interaction between minority status and type of schedule. These results were used for comparison purposes. Then scores on the posttest were randomly removed, at the 5 percent, 10 percent, 20 percent, and 25 percent level. Each of the four missing data techniques were applied once to the four different sets resulting in a total of 16 analyses. All four of the missing data techniques found similar effects with 10 percent missing data. At the 5 percent and 20 percent missing level, listwise did not detect the interaction found with the complete data set. The grand-mean technique failed to detect the interaction when 25 percent of the data were missing. In terms of the other statistics, for all levels of missing data, the cell-mean and simple regression techniques produced results similar to those of the actual data set.

The second study added the pretest to the Peabody to the data set from the first study which resulted in approximately 19 percent missing the pretest score. As before, each missing data technique was applied once to the expanded data set and statistics matching the first study were generated. All four missing data techniques produced similar results, from which the researchers concluded that one could feel confident about making conclusions, even with 19 percent of the pretest data missing.

Overall, grand-mean and cell-mean techniques produced the lowest standard deviations, and listwise deletion generally produced the highest. The simple regression and cell-mean techniques produced means most consistent with the actual means. Velotta issued a word of caution that even with only 5 percent missing data, listwise deletion failed to produce results matching the actual data. The recommendation was to use three or four missing data techniques even if the amount of missing data is small.

The listwise deletion, pairwise deletion, grand-mean substitution, and multiple regression missing data techniques were the objects of the study done by Roth, Campion, and Jones (1996). They investigated the impact of training success on job performance for 177 chemical process technicians. Tests and interviews were used to try to predict future performance. The data set containing the information on the technicians had an overall missing data rate of approximately 10 percent. The four missing data techniques were applied once to the data set and then regression equations were generated.

The researchers found that using different missing data techniques impacted findings. The multiple regression and grand-mean substitution approaches produced more significant regression coefficients, which was attributed to having a larger sample. Roth, Campion, and Jones recommended a decrease in the use of listwise deletion and an increase in the use of pairwise deletion and regression techniques because they preserve more information.

In a study which investigated changes in a two-variable prediction system, Hegamin-Younger and Forsyth (1998) investigated the effects of four imputation techniques. The listwise deletion, grand-mean substitution, cell-mean substitution, random hot-deck imputation, and 2-, 3-, and 4-variable regression missing data techniques were used to treat the data. Researchers were looking into how well ACT composite score and high school grade point average predicted college grade point average.

Nearly 19,000 students from 20 Missouri postsecondary schools were included in the sample. Listwise deletion was used to eliminate students not meeting the necessary data requirements to create a set used for comparison purposes. Student high school grade point averages were randomly removed until target proportions were met. Absolute deviation scores were used to compare regression coefficients generated from the comparison set and the treated sets. Mean absolute deviation scores were used to compare the college grade point averages predicted to the actual grade point averages across all students with a school and all schools.

The cell-mean substitution technique produced the smallest average deviation from the true high school grade point average regression coefficient and the grand-mean substitution technique produced the largest. The grand-mean substitution technique again produced the largest deviation from the true ACT composite score regression coefficient with the regression techniques producing the lowest deviations. In terms of college grade point average, the regression procedures produced the smallest deviation scores and the grand-mean substitution technique produced the largest deviation scores.

Hegamin-Younger and Forsyth drew three conclusions from their work. First, that the grand-mean substitution technique was not appropriate for handling missing data. Second, that the cell-mean substitution method should be used if the purpose is to estimate the regression coefficient of a variable with missing data. Lastly, if the purpose was to predict college grade point average, regression techniques should be used to handle missing data.

Discussion of Comparison Studies Using Actual Data

Because these studies utilized real data, researchers dealt with issues of designing studies with missing data in one of three ways: used a missing data technique to prepare the data for use, used the data as it was, or sought to use complete sets of data to work with. The listwise deletion technique was the only missing data technique used in the studies reviewed to prepare data for use.

The reviewed studies evaluated the performance of missing data techniques in one of two ways: comparison to a "true" set, or direct comparison of results produced by different missing data techniques. Comparisons made to a "true" set of data primarily used criteria in the form of deviation scores. The belief in these studies was that a "better" or "best" technique could be identified by finding the missing data technique with the smallest deviation scores. Studies without a "true" set use for comparison discussed similarities and differences in the performance of different missing data techniques. These discussions centered primarily on descriptive statistics and regression equation performance indicators.

Replication appeared to be of moderate concern. All removal of data was done in a random manner. The number of replications in the studies ranged from one to 250, with two studies using 100 or more, and the majority applying missing data techniques a single time.

Studies produced mixed performance of both listwise and pairwise deletion. Listwise deletion performed well in some studies (Kromrey & Hines, 1991, Thran & Gills, 1992; Ward & Clark, 1991) but its continued use for convenience reasons was not always supported (Raymond, 1987; Roth, Campion, & Jones, 1996). Concern for relying on pairwise deletion was also expressed (Raymond, 1987), yet it often performed well and was suggested for increased usage (Kromrey & Hines, 1991; Roth, Campion, & Jones, 1996; Ward & Clark, 1991).

Overall, the studies found grand-mean substitution to be the poorest performing (Kalton, 1983; Hegamin-Younger & Forsyth, 1998; Ward & Clark, 1991; Witta & Kaiser, 1991). Cell-mean substitution was found to be among top performers (Kalton, 1983; Hegamin-Younger & Forsyth, 1998; Velotta, 1995) as were regression techniques (Kalton, 1983; Roth, Campion, & Jones, 1996; Thran & Gillis, 1992), especially simple regression (Velotta, 1995; Ward & Clark, 1991).

Researchers sometimes found similarities in performance among missing data techniques investigated (Velotta, 1995; Ward & Clark, 1991; Witta & Kaiser, 1991), yet found that the choice of missing data techniques could impact the results of a study (Roth, Campion, & Jones, 1996; Ward & Clark, 1991). Recommendations from these studies suggest the use of multiple missing data techniques, to include simple regression and a zero-order technique (Velotta, 1995; Ward & Clark, 1991), and a comparison of results.

Table 2 presents a summary of the comparison studies that used real data. The table lists who and when the study was conducted, the missing data techniques investigated, and a brief listing of major results.

Table 2. Summary of comparison studies using real data

Author(s)	Methods compared	Results
Kalton (1983)	Grand-mean substitution Cell-mean substitution -8 and 10 cells Hot-deck imputation -8 and 10 cells Multiple regression Multiple regression + error	Cell-mean techniques and multiple regression perform best, grand-mean performs the worst, techniques with 10 cells outperform those with 8.
Raymond (1987)	Listwise deletion Pairwise deletion Iterative regression	No missing data technique is a reliable substitute for complete data, listwise and pairwise should not be relied on because of ease of use.
Ward & Clark (1991)	Listwise deletion Grand-mean substitution Simple regression Iterative regression	Choice of missing data technique can impact conclusions, impact of listwise deletion method not fully understood, grand-mean should not be used, suggest use of listwise and simple regression and compare results.
Witta & Kaiser (1991)	Listwise deletion Pairwise deletion Grand-mean substitution Multiple regression	Listwise, pairwise, and multiple regression equally effective in treating data to predict a criterion, grand-mean is not appropriate.
Kromrey & Hines (1991)	Listwise deletion Pairwise deletion Grand-mean substitution Simple regression Multiple regression	Effectiveness of missing data techniques should be evaluated against a criterion, under realistic situations, two deletion techniques performed better.
Thran & Gillis (1992)	Listwise deletion Grand-mean substitution Cell-mean substitution Hot-deck imputation -random and distance Multiple regression Multiple regression + error	Choice of best technique is not clear, hot-deck and regression + error techniques reasonable for statistics, listwise better at preserving relationships among variables.

Table 2. Continued

Author(s)	Methods compared	Results
Velotta (1995)	Listwise deletion Grand-mean substitution Cell-mean substitution Simple regression	Simulated study: Cell-mean and simple regression perform best. Actual data study: All techniques produced similar results. Three or four techniques should be applied even if missing data percent is small.
Roth, Campion, & Jones (1996)	Listwise deletion Pairwise deletion Grand-mean substitution Multiple regression	Choice of missing data technique can impact findings, consider decreased use of listwise and increase use of pairwise and regression to preserve data.
Hegamin-Younger & Forsyth (1998)	Listwise deletion Grand-mean substitution Cell-mean substitution Hot-deck imputation Multiple regression -2-, 3-, and 4-variable	Grand-mean is inappropriate, cell-mean should be used to estimate variable regression coefficient, regression procedures should be used to predict college grade point.

CHAPTER III. METHODOLOGY

This research was designed to investigate the impact of missing data techniques on indicators of student performance. The criterion-referenced tests used as the indicators of student achievement were developed as part of a three-year effort to renew and align K-12 curriculum and assessment. This effort was facilitated by the School Improvement Model project at Iowa State University. The use of these data and this study were approved by the Human Subjects Committee of Iowa State University.

Data

The data used in this study came from Lincoln County School District No. 1, in Kemmerer, Wyoming, with the permission of the district. The student performance data came from pretest scores obtained in the fall of 1995 and posttest scores obtained in the spring of 1996. Both student demographic data and achievement data in the areas of language arts, math, and reading from the 1995-1996 school year were used.

Language arts and math criterion-referenced test scores were available for all grades kindergarten through twelfth. Reading as a separate subject was tested for grades kindergarten through sixth. Statements referring to "all available subjects" will mean language arts, math, and reading for grades kindergarten through sixth; and language arts and math for grades seventh through twelfth.

To allow comparison across grade level, the scores on the criterion-referenced subject pretests and posttests were used to calculate a level of mastery score. This was done by dividing the score on the test by the number of possible, thus creating a "percent" correct indicator. The terminology, "level of mastery score," will be used throughout the remainder of this work.

Surrogate measures were used in two of the analyses. Representation of socioeconomic (SES) status was determined by enrollment in the district's free and reduced meal program. Students having missed 10 or more days from school were considered to be "high-absence" students.

The original Lincoln County data set

The original data set from Lincoln County contained criterion-referenced pretest and posttest information on 988 students in both math and language arts. The data set contained criterion-referenced pretests and posttests on 497 students in reading.

If the default approach of using listwise deletion was used, 12.9 percent of reading data, 15.2 percent of language arts data, and 22.9 percent of math data would have been lost. Tables 3, 4, and 5 present the percent of data lost using listwise deletion per grade level for reading, language arts, and math, respectively.

Table 3. Data lost using listwise deletion on reading test results

Grade level	Number of students	Number of students with both pre- and posttest scores	Percent lost
K	55	48	12.7
1	57	48	15.8
2	74	61	17.6
3	65	59	9.2
4	81	74	8.6
5	72	61	15.3
6	93	82	11.8
Total K-6	497	433	12.9

Table 4. Data lost using listwise deletion on language arts test results

Grade level	Number of students	Number of students with both pre- and posttest scores	Percent lost
K	55	48	12.7
1	57	37	35.1
2	74	61	17.6
3	65	61	6.2
4	81	74	8.6
5	72	57	20.8
6	93	78	16.1
7	77	75	2.6
8	84	75	10.7
9	89	82	7.9
10	88	73	17.0
11	80	55	31.3
12	73	62	15.1
Total K-12	988	838	15.2

Table 5. Data lost using listwise deletion on math test results

Grade level	Number of students	Number of students with both pre- and posttest scores	Percent lost
K	55	47	14.5
1	57	52	8.8
2	74	62	16.2
3	65	61	6.2
4	81	73	9.9
5	72	64	11.1
6	93	83	10.8
7	77	74	3.9
8	84	75	10.7
9	89	53	40.4
10	88	62	29.5
11	80	43	46.3
12	73	13	82.2
Total K-12	988	762	22.9

In order to conduct the analyses in part one of the study across all three subject areas, as well as create the complete set required for part two of the study, the listwise deletion missing data technique was applied across all three subjects. That is to say, when used, the listwise deletion method removed any student with a missing test score on any test in any subject. The end result was a smaller, yet complete, data set. Table 6 displays the data lost using this approach.

The data sets used

The following descriptions are used throughout the rest of this study:

ODS: The original data set from Lincoln County School District No. 1. There are data missing from this set. It is used to generate the comparison set in the second part of the study. In the

Table 6. Data lost using listwise deletion across all test results

Grade level	Number of students	Number of students with all pre- and posttest scores[a]	Percent lost
K	55	43	21.8
1	57	32	43.9
2	74	57	23.0
3	65	58	10.8
4	81	73	9.9
5	72	51	29.2
6	93	78	16.1
7	77	74	3.9
8	84	73	13.1
9	89	51	42.7
10	88	55	37.5
11	80	33	58.8
12	73	13	82.2
Total K-12	988	691	30.1

[a]Reading was only tested in grades K-6, and math and language arts were tested K-12.

first part of the study, five different missing data techniques are applied to the original data set.

CS: The comparison set. It is created from the ODS through the use of listwise deletion. It becomes the set used to compare against for the deviation measures. This set is the beginning set for all proportionally equivalent data sets that are created.

PEDs: The proportionally equivalent data sets. They were created from the CS by removing data in proportions that match missing data proportions in the ODS. They are treated as "smaller pictures" of the ODS. There were ten of these data sets, and each was treated by the five missing data techniques in the second part of the study. References to an individual PED will be done by using a hyphen and a number, i.e., PED-1, PED-2, etc.

Focus of the Results Investigated

Results in this study focused on two main indicators of student achievement. Student posttest level of mastery scores were investigated with the intent of providing information to assist in reviewing district impact on student achievement. Student gains in level of mastery were investigated with the intent of providing information to assist in reviewing the impact of classroom instruction on student achievement.

Study Design

This study was broken into two parts. The first part addressed the research question, "Would the use of different missing data techniques affect the results found with the data from Lincoln County School District No. 1?" The second part addressed the research question, "How well do the different missing data techniques perform?"

Common components

In order to allow for comparison of results, a uniform set of descriptive statistics and performance research questions were used in both parts of the study.

Descriptive statistics To reflect student achievement, average pretest, posttest, and gain score levels of mastery for each subject were generated by grade, gender, socioeconomic status, and level of absence. Correlations between pretest and posttest level of mastery scores for all subjects were also generated.

Performance research questions A total of nine questions was used to investigate student performance for a given set of data. The questions focused primarily on student performance disaggregated into demographic categories. The questions asked, hypothesis made, and analyses used were as follows:

Question 1: Did students' level of mastery improve from pretest to posttest?

Statement of Hypothesis: Level of mastery posttest scores will be significantly higher than pretest level of mastery scores in all subjects.

Statistical Procedure Used: For all but the data set created using the pairwise deletion missing data technique, a dependent t-test was used to compare posttest levels of mastery to posttest levels of mastery. When the pairwise deletion missing data technique was used, the sets of students with pretest scores did not match the set of students with posttest scores. Thus an independent t-test was used.

Question 2: Do gains in level of mastery differ between males and females?

Statement of Hypothesis: There will be no difference in gain in level of mastery between males and females in all subjects.

Statistical Procedure Used: An independent t-test was used to compare the performance of males to females. This was done in all subjects.

Question 3: Do gains in level of mastery differ between high- and low-SES students?

Statement of Hypothesis: There will be no difference in gain in level of mastery between high- and low-SES students in all subjects.

Statistical Procedure Used: An independent t-test was used to compare the performance of high-SES student to low-SES students. This was done in all subjects.

Question 4: Do gains in level of mastery differ between students with high levels of absence and low levels of absence?

Statement of Hypothesis: There will be no difference in gain in mastery between students with high levels of absence and low levels of absence.

Statistical Procedure Used: An independent t-test was used to compare the performance of high absence students to low absence students. This was done in all subjects.

Question 5: Do gains in level of mastery differ by grade level?

Statement of Hypothesis: There will be no difference in gain in mastery between grade levels.

Statistical Procedure Used: A one-way analysis of variance (ANOVA) was used to compare the performance of students by grade level. This was done in all subjects.

Question 6: Do posttest levels of mastery differ between males and females?

Statement of Hypothesis: There will be no difference in posttest levels of mastery between males and females in all subjects.

Statistical Procedure Used: An independent t-test was used to compare the posttest levels of mastery of males to females. This was done in all subjects.

Question 7: Do posttest levels of mastery differ between high- and low-SES students?

Statement of Hypothesis: There will be no difference in posttest levels of mastery between high- and low-SES students in all subjects.

Statistical Procedure Used: An independent t-test was used to compare the posttest levels of mastery of high-SES student to low-SES students. This was done in all subjects.

Question 8: Do posttest levels of mastery differ between students with high levels of absence and low levels of absence?

Statement of Hypothesis: There will be no difference in posttest levels of mastery between students with high levels of absence and low levels of absence.

Statistical Procedure Used: An independent t-test was used to compare the posttest levels of mastery of high absence students to low absence students. This was done in all subjects.

Question 9: Do posttest levels of mastery differ by grade level?

Statement of Hypothesis: There will be no difference in posttest levels of mastery between grade levels.

Statistical Procedure Used: A one-way analysis of variance (ANOVA) was used to compare the posttest levels of mastery of students by grade level. This was done in all subjects.

Part 1 – Impact on the original data set

The research question addressed in this part of the study was, "Would the use of different missing data techniques affect the results found with the data from Lincoln County School District

No. 1?" The original data set (ODS) from Lincoln County School District No. 1 was used in this part of the study.

The ODS was treated with five different missing data techniques. This created five distinct data sets for which the standard set of descriptive statistics was generated. In addition, the statistical procedures required to address the nine performance research questions were also performed on the five distinct data sets. Because there was no "true" set of values to use for comparison, the results were presented, by question, with results from all five missing data techniques displayed.

Part 2 - Performance of the missing data techniques

The research question addressed in this part of the study was, "How well do the different missing data techniques perform?" The difficulty in answering this question was that the data were incomplete. To measure the performance of the missing data techniques required some indicator or criterion by which the techniques could be compared. This part of the study was structured as a simulation to address this issue.

Instead of applying the five missing data techniques to the original student performance data, simulated data sets were created that allowed for a comparison of performance of the missing data techniques. The proportionally equivalent data sets were created to simulate patterns that existed in the original data set. Though the process to create the proportionally equivalent data set removed data randomly, the process was repeated ten times to negate the impact of an odd removal pattern. The descriptive statistics and deviation measures for each missing data technique were averaged across the ten proportionally equivalent data sets.

Proportions of missing data The proportions of missing pretest and posttest data were calculated using criterion-referenced tests scores for all subjects, for each grade level, grades kindergarten through twelfth. The proportion of missing data, expressed as a decimal, was calculated by dividing the number of students for whom a score was missing by the total number of students examined. Within each subject, proportions of students missing pretest, posttest, or both scores were calculated. These proportions were used to create data sets with missing data from the sample used in this study.

Proportionally equivalent data sets Using the proportions of missing data, the number of pretest and posttest scores to be removed to create a proportionally equivalent data set was determined. This was done by calculating the product of the proportion, expressed as a decimal, and the number of students in each subgroup. The calculated number of pretest and posttest scores were removed by random.

To alleviate problems associated with the removal of a particularly misrepresentative group of scores, the process was repeated ten times, similar to the method used by Kalton (1983). This produced ten proportionally equivalent data sets to which five missing data techniques were applied. To allow comparison of results, the five missing data techniques were applied to the same ten proportionally equivalent data sets. Both the descriptive statistics used for comparison of the five missing data techniques and the deviation measures used to evaluate the quality of the imputation missing data techniques were averaged across the ten proportionally equivalent data sets.

Deviation measures The second part of the study evaluated the performance of the missing data techniques through the use of deviation measures. For the comparison set and the results of the five missing data techniques, the mean and standard deviation for pretest, posttest, and gain scores will be calculated. Because the calculation of the gain score mean for the pairwise deletion missing data technique was done directly by calculating the difference between the posttest mean and the pretest mean, a standard deviation for the gain scores for this technique was not available.

This first required the calculation and reporting of the standard descriptive statistics for each of the ten proportionally equivalent data sets (PEDs) and the comparison set (CS). Deviation measures were then calculated using the statistics from the CS as the "true" values and the statistics from the PEDs as the estimated values.

The deviation measures were used as the criteria for evaluating missing data technique performance. The mean deviation score was used as an indicator of bias in imputed values and the mean absolute deviation score was used to indicate how "close" imputed values were to actual values. Missing data technique performance was seen as improving as deviation measure scores decreased. In other words, the smaller the better.

The process The original data set (ODS) was turned into the comparison set (CS) through the use of listwise deletion. The CS was a set of data, complete and rectangular in nature, with no missing data. Examining the data that was removed, missing data proportions (MDPs) were calculated. These MDPs were used to create a proportionally equivalent data set (PED) by randomly removing data, a process that was repeated ten times. Each of the ten PEDs was treated by the five missing data techniques.

The descriptive data for each missing data technique were averaged over the ten PEDs and reported with the descriptive data for the CS. The deviation statistics for examining the performance of each imputation technique was also averaged over the ten PEDs. In all cases, values from the CS as well as the statistics generated from the CS were seen as the "true" values.

Analysis of Data

All data in this study were imported into the statistical software package SPSS, which was used to conduct all analyses. For each set of data examined, descriptive statistics were generated first. These statistics included means and correlations.

Specific statistical tests were performed to address the performance research questions. Statistical procedures conducted were one-way analysis of variance (ANOVA), independent t-tests, and dependent t-tests.

In the second part of the study, deviation scores were calculated to compare performance of the missing data techniques.

Treatment of Data

In this study, data were treated three ways: the application of missing data techniques, the conduction of statistical procedures, and the calculation of deviation measures.

Missing data techniques

A total of five different missing data techniques were applied to the original Lincoln County data set. The techniques were only applied to pretest and posttest level of mastery scores. No attempt was made to recover missing demographic information in the data set. The missing data techniques used were:

1. Listwise deletion. Any student with data missing in any of the variables of interest was removed from the data set. With the exception of creating the complete data set for part two of the study, this was done subject by subject. The result was a smaller, but complete, rectangular set of data.

2. Pairwise deletion. The data were left as they were, with any analysis performed using only available data. The result was a data set with differing levels of missing data for the variables in the data set. It was known that this technique sometimes resulted in the sets and number of students used in varying analyses to be different.
3. Grand-mean imputation. For each subject, pretest and posttest level of mastery averages were calculated on available scores across all grade levels. These averages were then substituted in for appropriate missing values. The result was a complete, rectangular set of data.
4. Cell-mean imputation. For each subject, pretest and posttest level of mastery means were calculated for each grade using available data. These means were then substituted in for appropriate grade-level missing values. The result was a complete, rectangular set of data.
5. Simple regression. Correlations were calculated between all available subject pretest and posttest scores. These correlations were calculated using data from students possessing scores for all tests in all subjects. Because reading tests were only administered through grade six, one set of correlations was generated for use in grades kindergarten through sixth and another set of correlations was generated for grades seven through 12. By subject, missing values for each student were replaced with a value generated from a regression equation created using an available test score from the same student with the highest correlation to the missing value. When test scores were not available for creating a regression equation, the grade and subject appropriate pretest average level of mastery was substituted in for the missing pretest value and the posttest missing value replaced with the resulting regression estimate. The result was a complete, regular set of data.

Statistical procedures

In determining the significance level for the statistical procedures performed in this study, alpha was set at 0.05. Thus, the value of $p=0.05$ was used as an acceptable level of type I error, or the chance of rejecting the null hypothesis when it is actually true. This level was selected because "p usually is set at 0.05" (Gall, Borg, & Gall, 1996, p. 187).

The t-test is used to examine the statistical significance of an observed difference between two sample means (Gall, Borg, & Gall, 1996). A dependent t-test is used when "...the scores under one condition are dependent on the scores in the other condition" (Hinkle, Wiersma, & Jurs, 1988, p. 253). Such was the case in Question 1 when matched pretest and posttest scores were used to compare pretest and posttest levels of mastery. The formula for the t-test is:

$$t = \frac{X1 - X2}{S_{x1-x2}}$$

Independent t-tests were used to examine differences in gain in level of mastery and posttest levels of mastery disaggregated by gender, socioeconomic status, and level of attendance. An independent t-test was also used to examine the difference in pretest and posttest levels of mastery in data sets created using the pairwise deletion missing data technique.

One-way analysis of variance (ANOVA) procedures were used to examine differences in gains in level of mastery and posttest levels of mastery disaggregated by grade level. The ANOVA was designed to "...test the hypothesis of equality of K population means while maintaining the Type I error rate at the preestablished levels for the entire set of comparisons" (Hinkle, Wiersma, & Jurs, 1988, p. 329).

The statistic used in ANOVA is the F ratio, which is the ratio of two variance estimates: the mean square between groups and the mean square within groups. The formula for the analysis of variance is:

$$F = \frac{MS_b}{MS_w} = \frac{SS_b/df_b}{SS_w/df_w}$$

The Scheffé post hoc comparison test was used to determine which group means differed significantly.

Deviation measures

The simulation created in the second part of this study allowed for the use of deviation measures to judge the performance of the missing data techniques. Two deviation measures were used to evaluate the missing data techniques: the mean deviation and the mean absolute deviation.

The mean deviation was calculated by summing the differences between all actual values and all estimated, or imputed, values and dividing by the number of differences. Because of its ability to indicate, on average, whether the imputed values were above or below the actual values, the mean deviation was used as a measure of bias in the imputed values (Kalton, 1983).

The mean absolute deviation was calculated by summing the absolute value of the differences between all actual values and all estimated, or imputed, values and dividing by the number of deviations. Because it measures the average "distance" between the imputed values and the actual values, the mean absolute deviation was used to measure the degree of "closeness" in the imputed values (Kalton, 1983).

CHAPTER IV. FINDINGS

The efforts of this study were divided into two parts: 1) the impact of missing data technique selection on student performance results, and 2) the performance of the missing data techniques used in this study to represent student performance data.

Part 1

The primary problem of this study was to determine whether the selection of a missing data technique to handle missing data impacted the indicators of student achievement. The data were criterion-referenced, pretest and posttest scores in reading, mathematics, and language arts from a single academic year. Testing in reading was done in grades K-6, while testing in mathematics and language arts was done in grades K-12.

In order to allow easier comparison, test scores were divided by the number of items on the tests which produced a "percent correct" indicator. In this study, this percent is referred to as a "level of mastery."

The investigation of the impact of missing data technique selection focused on the results tied to nine research questions in each of the three academic subjects. This portion of the chapter will report on descriptive statistics and then results, by subject, on the nine research questions.

In data treated by the pairwise deletion missing data technique, gains in level of mastery were calculated directly by subtracting pretest scores from posttest scores. Though this resulted in a gain score, no distribution of scores exists upon which to run statistical tests. Thus, sets treated with pairwise deletion will not have statistics to compare when investigating differences in gain scores.

Descriptive statistics – Reading

Descriptive statistics on student performance in reading in the form of pretest, posttest, and gain in mastery scores are given in Tables 7-9. Across the five missing data techniques, means on the tests were close. The differences between the highest and lowest means were 0.38, 0.42, and 0.38 percent for pretest, posttest, and gains in level of mastery scores, respectively. Standard deviations generated across the five missing data techniques were also close. The differences between the highest and lowest standard deviation were 0.85, 1.06, and 1.07 for pretest, posttest, and gains in mastery scores, respectively.

Descriptive statistics - Math

Descriptive statistics on student performance in reading in the form of pretest, posttest and gain in mastery scores are given in Tables 10–12. Across the five missing data techniques, the

Table 7. Mean percents and standard deviations in criterion-referenced, K-6 reading pretest level of mastery scores,[a] across all grades, by missing data technique

MDT	Cases	Mean	S.D.
Listwise	392	60.92	17.62
Pairwise	451	60.77	17.95
Grand mean	497	60.77	17.10
Cell mean	497	60.88	17.32
Simple regression	497	61.15	17.45

[a]Level of mastery score = percent correct.

Table 8. Mean percents and standard deviations in criterion-referenced, K-6 reading posttest level of mastery scores, across all grades, by missing data technique

MDT	Cases	Mean	S.D.
Listwise	392	73.54	18.62
Pairwise	469	73.21	19.68
Grand mean	497	73.21	19.12
Cell mean	497	73.12	19.43
Simple regression	497	73.44	19.27

Table 9. Mean percents and standard deviations in gains in criterion-referenced, K-6 reading level of mastery scores, across all grades, by missing data technique

MDT	Cases	Mean	S.D.
Listwise	392	12.62	15.40
Pairwise[a]	–	12.44	–
Grand mean	497	12.44	16.47
Cell mean	497	12.24	15.93
Simple regression	497	12.29	15.59

[a]Gains obtained through direct calculation since no distribution of scores exists.

Table 10. Mean percents and standard deviations in criterion-referenced, K-12 math pretest level of mastery scores, across all grades, by missing data technique

MDT	Cases	Mean	S.D.
Listwise	691	47.02	13.53
Pairwise	815	46.82	13.60
Grand mean	988	46.82	12.35
Cell mean	988	46.34	12.37
Simple regression	988	46.39	12.67

Table 11. Mean percents and standard deviations in criterion-referenced, K-12 math posttest level of mastery scores, across all grades, by missing data technique

MDT	Cases	Mean	S.D.
Listwise	691	64.54	18.34
Pairwise	828	64.67	18.94
Grand mean	988	64.67	17.34
Cell mean	988	63.80	17.63
Simple regression	988	63.34	17.99

Table 12. Mean percents and standard deviations in gains in criterion-referenced, K-12 math level of mastery scores, across all grades, by missing data technique

MDT	Cases	Mean	S.D.
Listwise	691	17.47	16.07
Pairwise[a]	--	17.85	--
Grand mean	988	17.85	15.77
Cell mean	988	17.46	15.74
Simple regression	988	16.95	15.44

[a]Gains obtained through direct calculation since no distribution of scores exists.

differences in means on the math tests were larger than those on the reading tests. The differences between the highest and lowest means were 0.68, 1.33, and 0.90 percent for pretest, posttest, and gains in level of mastery scores, respectively. The differences between the highest and lowest standard deviation were 1.25, 1.60, and 0.63 for pretest, posttest, and gains in mastery scores, respectively.

Descriptive statistics - Language Arts

Descriptive statistics on student performance in reading in the form of pretest, posttest, and gain in mastery scores are given in Tables 13-15. Across the five missing data techniques, the differences in means on the language arts tests were among the largest across all tests and subjects. The differences between the highest and lowest means were 2.75, 3.26, and 0.57 percent for pretest, posttest, and gains in level of mastery scores, respectively. The differences

Table 13. Mean percents and standard deviations in criterion-referenced, K–12 language arts pretest level of mastery scores, across all grades, by missing data technique

MDT	Cases	Mean	S.D.
Listwise	691	54.16	19.50
Pairwise	901	51.41	19.47
Grand mean	988	51.41	18.59
Cell mean	988	51.44	19.14
Simple regression	988	51.53	19.03

Table 14. Mean percents and standard deviations in criterion-referenced, K–12 language arts posttest level of mastery scores, across all grades, by missing data technique

MDT	Cases	Mean	S.D.
Listwise	691	65.35	20.32
Pairwise	907	62.44	20.95
Grand mean	988	62.44	20.07
Cell mean	988	62.09	20.72
Simple regression	988	62.12	20.49

Table 15. Mean percents and standard deviations in gains in criterion-reference, K-12 language arts level of mastery scores, across all grades, by missing data technique

MDT	Cases	Mean	S.D.
Listwise	691	11.16	16.70
Pairwise[a]	--	11.03	--
Grand mean	988	11.03	17.06
Cell mean	988	10.66	17.24
Simple regression	988	10.59	15.41

[a]Gains obtained through direct calculation since no distribution of scores exists.

between the highest and lowest standard deviation were 0.91, 0.88, and 1.83 for pretest, posttest, and gains in mastery scores, respectively.

Descriptive statistics - Correlations

To monitor possible impact on relationships between scores, correlations between criterion-referenced pretest and posttest scores were generated among all subjects. These are shown in Tables 16-20.

Table 16 displays the correlations generated from data treated with the listwise deletion missing data technique. All of the correlations were statistically significant ($p<.001$). The two largest correlations existed between reading posttest scores and math posttest scores ($r=.656$) and language arts pretest and posttest scores ($r=.648$). The smallest correlation existed between reading posttest scores and language arts pretest scores ($r=.200$). Average subject test correlations, using both pretests and posttests, showed that reading correlations were the largest

Table 16. Intercorrelations among criterion-referenced, pretest and posttest level of mastery scores, across all grades, from data treated with the listwise deletion missing data technique

Variable	Reading pretest	Reading posttest	Math pretest	Math posttest	LA pretest	LA posttest
Reading pretest	--	.640	.634	.550	.477	.641
Reading posttest	--	--	.381	.656	.200	.610
Math pretest	--	--	--	.527	.429	.421
Math posttest	--	--	--	--	.364	.607
LA pretest	--	--	--	--	--	.648
LA posttest	--	--	--	--	--	--

Note. $p < .001$ for all cases; $n=392$ for correlations involving reading tests scores, $n=691$ for all others.

($r=.549$), followed by math ($r=.510$) and language arts ($r=.505$). This may be connected to a need read well in order to perform well on written tests.

Correlations generated from data treated with pairwise deletion are given in Table 17. All of the correlations generated were statistically significant ($p < .001$). Correlations ranged from a high of $r=.692$ between reading posttest scores and math posttest scores, to a low of $r=.178$ between reading posttest and language arts pretest scores. Average subject test correlations, using both pretests and posttests, showed that reading correlations were the largest ($r=.542$), followed by language arts ($r=.506$) and math ($r=.504$). This may be connected to a need read well in order to perform well on written tests.

Table 17. Intercorrelations among criterion-referenced, pretest and posttest level of mastery scores, across all grades, from data treated with the pairwise deletion missing data technique

Variable	Reading pretest	Reading posttest	Math pretest	Math posttest	LA pretest	LA posttest
Reading pretest	--	.640 (433)	.552 (443)	.574 (439)	.463 (435)	.651 (420)
Reading posttest	--	--	.390 (435)	.692 (463)	.178 (428)	.643 (446)
Math pretest	--	--	--	.524 (763)	.414 (793)	.413 (765)
Math posttest	--	--	--	--	.339 (772)	.614 (785)
LA pretest	--	--	--	--	--	.670 (840)
LA posttest	--	--	--	--	--	--

Note. Number of pairs used to calculate correlation is given in parentheses; $p < =.001$ for all cases.

All correlations generated between test scores in data treated with the grand-mean substitution missing data technique were statistically significant ($p < .001$). These are listed in Table 18. Correlations ranged from a high of $r = .679$ between reading posttest scores and math posttest scores, to a low of $r = .166$ between reading posttest and language arts pretest scores. Average subject test correlations, using both pretests and posttests, showed that reading correlations were the largest ($r = .510$), followed by math ($r = .466$) and language arts ($r = .463$). This may be connected to a need to read well in order to perform well on written tests.

Table 18. Intercorrelations among criterion-referenced, pretest and posttest level of mastery scores, across all grades, from data treated with the grand-mean substitution missing data technique

Variable	Reading pretest	Reading posttest	Math pretest	Math posttest	LA pretest	LA posttest
Reading pretest	--	.591	.540	.533	.449	.588
Reading posttest	--	--	.359	.679	.166	.604
Math pretest	--	--	--	.477	.388	.364
Math posttest	--	--	--	--	.303	.539
LA pretest	--	--	--	--	--	.613
LA posttest	--	--	--	--	--	--

Note. n=497 for correlations involving reading tests scores, n=988 for all others; $p<=.001$ for all cases.

Table 19 displays the correlations generated from data treated with the cell-mean substitution missing data technique. All of the correlations were statistically significant ($p<.001$). The largest correlation existed between reading posttest scores and math posttest scores ($r=.686$). The smallest correlation existed between reading posttest scores and language arts pretest scores ($r=.122$). Average subject test correlations, using both pretests and posttests, showed that reading correlations were the largest ($r=.521$), followed by math ($r=.485$) and language arts ($r=.476$). This may be connected to a need read well in order to perform well on written tests.

Correlations generated from data treated with the simple regression missing data technique are given in Table 20. All of the correlations generated were statistically significant ($p<.001$). Unlike the other data sets, the largest correlation of $r=.698$ existed between language arts pretest

Table 19. Intercorrelations among criterion-referenced, pretest and posttest level of mastery scores, across all grades, from data treated with the cell-mean substitution missing data technique

Variable	Reading pretest	Reading posttest	Math pretest	Math posttest	LA pretest	LA posttest
Reading pretest	--	.630	.541	.561	.436	.619
Reading posttest	--	--	.361	.686	.122	.619
Math pretest	--	--	--	.499	.412	.390
Math posttest	--	--	--	--	.327	.577
LA pretest	--	--	--	--	--	.628
LA posttest	--	--	--	--	--	--

Note. n=497 for correlations involving reading tests scores, n=988 for all others; p<=.001 for all cases.

Table 20. Intercorrelations among criterion-referenced, pretest and posttest level of mastery scores, across all grades, from data treated with the simple regression missing data technique

Variable	Reading pretest	Reading posttest	Math pretest	Math posttest	LA pretest	LA posttest
Reading pretest	--	.644	.559	.569	.484	.633
Reading posttest	--	--	.409	.695	.228	.589
Math pretest	--	--	--	.539	.441	.427
Math posttest	--	--	--	--	.390	.614
LA pretest	--	--	--	--	--	.698
LA posttest	--	--	--	--	--	--

Note. n=497 for correlations involving reading tests scores, n=988 for all others; p<=.001 for all cases.

and posttest scores. Similar to the other four data sets, the smallest correlation (r=.228) existed between reading posttest and language arts pretest scores. Average subject test correlations, using both pretests and posttests, showed that reading correlations were the largest (r=.545), followed by language arts (r=.520) and math (r=.518). This may be connected to a need read well in order to perform well on written tests.

Of the 15 possible correlations between pretest and posttest scores, those generated from data treated with the grand-mean substitution missing data technique were the smallest 11 times and the second smallest four times. Data treated with the cell-mean substitution appeared to generate the next smallest set of correlations with two of the smallest and ten of the second smallest values. Correlations generated from data treated with the simple regression missing data technique produced the largest correlations 11 times and the second largest twice. Correlations generated from data treated with listwise and pairwise deletion missing data techniques tended to fall within the extremes previously mentioned.

Research questions - Reading

To study the impact that missing data technique selection might have, nine research questions were developed. This section reports on the findings for reading.

Research Question 1 - Pretest to posttest The first research question asked whether students' level of mastery improved from pretest to posttest. The first research hypothesis stated that posttest level of mastery scores would be significantly higher than pretest level of mastery scores. An independent t-test was used to test this hypothesis in data treated with the pairwise deletion missing data technique. Dependent t-tests were used to test the hypothesis in data treated with the other four missing data techniques.

Table 21 shows the mean, standard deviation, and test results for student performance on criterion-referenced reading tests across all five missing data techniques. The null hypothesis that there would be no significant difference in pretest and posttest level of mastery scores in reading was rejected across all five missing data techniques ($t=16.229$, $p=.000$; $t=10.019$, $p=.000$; $t=16.820$, $p=.000$; $t=17.129$, $p=.000$; $t=17.575$, $p=.000$). It can be concluded that, across data treated with all five missing data techniques, posttest level of mastery scores on the criterion-referenced reading tests are significantly higher than the pretest level of mastery scores on the criterion-referenced reading tests. This indicates, across all five data sets, that the students in this

Table 21. Analysis of criterion-referenced, mean K-6 reading pretest and posttest level of mastery scores, across all grades, by missing data technique

MDT	Test	Number of pairs	Mean	S.D.	t-value
Listwise					
	Pretest	392	60.92	17.62	16.229**
	Posttest	392	73.54	18.62	
Pairwise					
	Pretest	451	60.77	17.95	10.019**
	Posttest	469	73.21	19.68	
Grand mean					
	Pretest	497	60.77	17.10	16.82**
	Posttest	497	73.21	19.12	
Cell mean					
	Pretest	497	60.88	17.32	17.129**
	Posttest	497	73.12	19.43	
Simple regression					
	Pretest	497	61.15	17.45	17.575**
	Posttest	497	73.44	19.27	

**$p<.000$.

study significantly improved their performance in reading. These results support that actual learning occurred with these students in the subject of reading.

Research Question 2 - Gain by gender The second research question asked whether students' gains in level of mastery on criterion-referenced reading tests would differ between males and females. The second research hypothesis stated that there would be no significant difference in gain in level of mastery between males and females. Independent t-tests were used to test this hypothesis.

Because gains in mastery in data treated with the pairwise deletion missing data technique were generated by subtracting the pretest mean from the posttest mean which did not create a distribution of scores, tests to check for significance of differences could not be performed. This resulted in reporting statistics on only four of the five missing data techniques on research questions addressing gains in level of mastery.

Table 22 shows the mean, standard deviation, and test results for student performance on criterion-referenced reading tests across four of the five missing data techniques. The hypothesis that there would be no significant difference in gain in level of mastery scores between males and females was not rejected ($t=0.294$; $t=0.047$; $t=0.226$; $t=0.386$). Males and females did not demonstrate significantly different gains in mastery on criterion-referenced reading tests in data treated with four of the five missing data techniques. Females did slightly outperform males in each of the data sets. The results support that, in terms of reading, males and females were learning equally well during the year.

Research Question 3 - Gain by SES The third research question asked whether students' gain in level of mastery on criterion-referenced reading tests would differ between high and low socioeconomic (SES) students. The third research hypothesis stated that there would be

Table 22. Analysis of male and female gains in criterion-referenced, K-6 reading level of mastery scores, across all grades, by missing data technique

MDT	Gender	Number of pairs	Mean	S.D.	t-value
Listwise					
	Female	193	12.85	15.67	0.294
	Male	199	12.40	15.17	
Pairwise[a]					
	Female	--	12.46	--	
	Male	--	12.42	--	
Grand mean					
	Female	242	12.47	16.32	0.047
	Male	255	12.40	16.65	
Cell mean					
	Female	242	12.40	16.37	0.226
	Male	255	12.08	15.52	
Simple regression					
	Female	242	12.57	15.61	0.386
	Male	255	12.03	15.60	

[a]Gains obtained through direct calculation since no distribution of scores exists.

no significant difference in gain in level of mastery between high- and low-SES students. Independent t-tests were used to test this hypothesis.

Table 23 shows the mean, standard deviations, and test results for student performance on criterion-referenced reading tests across four of the five missing data techniques. Because gains in mastery in data treated with the pairwise deletion missing data technique were generated by subtracting the pretest mean from the posttest mean which did not create a distribution of scores, a t-test could not be performed.

The hypothesis that there would be no significant difference in gain in level of mastery scores between high- and low-SES students was rejected in data treated with the listwise deletion, grand-mean substitution, and cell-mean substitution missing data techniques ($t=-2.228$, $p<.05$; $t=-2.450$, $p<.05$; $t=2.001$, $p<.05$). In these data sets, it can be concluded that low-SES students achieve significantly higher gains in mastery scores on the criterion-referenced reading tests than did high-SES students.

Table 23. Analysis of low socioeconomic and middle and high socioeconomic student gains in criterion-referenced, K-6 reading level of mastery scores, across all grades, by missing data technique

MDT	SES	Number of pairs	Mean	S.D.	t-value
Listwise					
	Middle and high	344	11.98	15.04	-2.228*
	Low	48	17.24	17.25	
Pairwise[a]					
	Middle and high	--	11.64	--	
	Low	--	18.21	--	
Grand mean					
	Middle and high	427	11.71	16.14	-2.450*
	Low	70	16.88	17.83	
Cell mean					
	Middle and high	427	11.66	15.67	-2.001*
	Low	70	15.76	16.98	
Simple regression					
	Middle and high	427	11.77	15.34	-1.855
	Low	70	15.49	16.78	

[a]Gains obtained through direct calculation since no distribution of scores exists.

*$p<.05$.

The hypothesis that there would be no significant difference in gain in level of mastery scores between high- and low-SES students was not rejected in data treated with the simple regression missing data technique (t=-1.855). In this data set, high- and low-SES students did not demonstrate significantly different gains in mastery on the criterion-referenced reading tests.

Whether statistically significant or not, low-SES students gained more than the middle and high-SES students in all five data sets. These results support that low-SES students made more progress in the area of reading than did other students.

Research Question 4 - Gain by level of absence The fourth research question asked whether students' gains in level of mastery on criterion-referenced reading tests would differ by level of absence. The fourth research hypothesis stated that there would be no significant difference in gain in level of mastery between students with high levels of absence and normal levels of absence. Independent t-tests were used to test this hypothesis.

Table 24 shows the mean, standard deviation, and test results for student performance on criterion-referenced reading tests across four of the five missing data techniques. Because gains in mastery in data treated with the pairwise deletion missing data technique were generated by subtracting the pretest mean from the posttest mean which did not create a distribution of scores, a t-test could not be performed. The hypothesis that there would be no significant difference in gain in level of mastery scores between students with high levels of absence and normal levels of absence was not rejected (t=-0.848; t=0.249; t=0.083; t=0.196). Students with high levels of absence and normal levels of absence did not demonstrate significantly different gains in mastery on criterion-referenced reading tests in data treated with four of the five missing data techniques.

Table 24. Analysis of high absence and normal attendance student gains in criterion-referenced, K-6 reading level of mastery scores, across all grades, by missing data technique

MDT	Absence	Number of pairs	Mean	S.D.	t-value
Listwise					
	Normal	306	12.22	14.44	-0.848
	High	86	14.05	18.44	
Pairwise[a]					
	Normal	--	12.53	--	
	High	--	12.11	--	
Grand mean					
	Normal	369	12.56	15.48	0.249
	High	128	12.09	19.11	
Cell mean					
	Normal	369	12.28	15.05	0.083
	High	128	12.13	18.30	
Simple regression					
	Normal	369	12.38	14.68	0.196
	High	128	12.03	18.01	

[a]Gains obtained through direct calculation since no distribution of scores exists.

Research Question 5 – Gain by grade The fifth research question asked whether students' gains in level of mastery differed by grade. The fifth research hypothesis stated that there would be no significant difference in gain in level of mastery between grade levels. A one-way analysis of variance (ANOVA) was used to test this hypothesis.

Tables 25–29 show the means and available standard deviations for student performance on the criterion-referenced reading tests across the five missing data techniques. Across all five missing data techniques, the fifth grade was the only one to report a negative change. Again

Table 25. Mean percents and standard deviations for gains in criterion-referenced, K-6 reading level of mastery scores, by grade, from data treated with the listwise deletion missing data technique

Grade	Cases	Mean	S.D.
K	43	24.44	16.40
1	32	26.94	12.03
2	57	16.25	15.63
3	58	14.09	11.77
4	73	7.64	8.56
5	51	-2.21	15.20
6	78	10.85	12.36

Table 26. Mean percents for gains in criterion-referenced, K-6 reading level of mastery scores, by grade, from data treated with the pairwise[a] deletion missing data technique

Grade	Cases	Mean	S.D.
K	--	24.09	--
1	--	25.78	--
2	--	15.65	--
3	--	14.21	--
4	--	7.54	--
5	--	-3.71	--
6	--	9.28	--

[a]Gains obtained through direct calculation since no distribution of scores exists.

Table 27. Mean percents and standard deviations for gains in criterion-referenced, K–6 reading level of mastery scores, by grade, from data treated with the grand-mean substitution missing data technique

Grade	Cases	Mean	S.D.
K	55	23.47	16.27
1	57	25.25	13.06
2	74	16.36	15.92
3	65	14.64	11.37
4	81	7.84	8.64
5	72	-2.47	16.99
6	93	8.93	15.15

Table 28. Mean percents and standard deviations for gains in criterion-referenced, K–6 reading level of mastery scores, by grade, from data treated with the cell-mean substitution missing data technique

Grade	Cases	Mean	S.D.
K	55	24.09	15.53
1	57	25.78	13.26
2	74	15.65	14.61
3	65	14.21	11.20
4	81	7.54	8.58
5	72	-3.71	15.76
6	93	9.28	13.04

Table 29. Mean percents and standard deviations for gains in criterion-referenced, K-6 reading level of mastery scores, by grade, from data treated with the simple regression missing data technique

Grade	Cases	Mean	S.D.
K	55	23.37	15.62
1	57	24.51	12.03
2	74	15.46	14.61
3	65	13.82	11.24
4	81	8.26	8.53
5	72	-2.25	16.11
6	93	9.43	13.95

across all five missing data techniques, the order of gain scores for the grades, from highest to lowest, was: first, kindergarten, second, third, sixth, fourth, and fifth.

Table 30 shows the results of the ANOVA of student performance by grade level. Given the interest in determining if the choice of missing data technique impacted the results of the ANOVA, the results of the Scheffé multiple range tests for differences between specific pairs of grades are not reported here. Scheffé multiple range tests were run and can be found in Appendix B.

The null hypothesis that would be no significant difference in gain in level of mastery between grade levels was rejected across all data sets ($F=25.971$, $p=.000$; $F=30.241$, $p=.000$; $F=38.363$, $p=.000$; $F=31.374$, $p=.000$). It can be concluded that, across data treated with four of the five missing data techniques, gains in level of mastery differed significantly by grade level.

Research Question 6 – Posttest by gender The sixth research question asked whether student posttest level of mastery scores on criterion-referenced reading tests would differ between males and females. The hypothesis stated that there would be no significant difference in posttest

Table 30. One-way analysis of variance of gain in criterion referenced, K-6 reading level of mastery scores, by grade, for all missing data techniques[a]

MDT	Source	df	Sum of squares	Mean squares	F ratio
Listwise					
	Between groups	6	2.671	0.445	25.971***
	Within groups	385	6.599	0.017	
	Total	391	9.270		
Grand mean					
	Between groups	6	3.636	0.606	30.241***
	Within groups	490	9.820	0.020	
	Total	496	13.457		
Cell mean					
	Between groups	6	4.022	0.670	38.363***
	Within groups	490	8.562	0.017	
	Total	496	12.584		
Regression					
	Between groups	6	3.346	0.558	31.374***
	Within groups	490	8.709	0.018	
	Total	496	12.054		

[a]Pairwise deletion did not produce a distribution of scores and was excluded.

***$p < .001$.

level of mastery scores between males and females. This hypothesis was tested using an independent t-test.

Table 31 displays the mean, standard deviation, and posttest results for student performance on criterion-referenced reading tests across the five missing data techniques. The hypothesis that there would be no significant difference in posttest level of mastery scores between males and females was not rejected (t=0.422; t=0.547; t=0.547; t=0.593; t=0.614). Males and females

Table 31. Analysis of male and female, criterion-referenced, K–6 reading posttest level of mastery scores, across all grades, by missing data technique

MDT	Gender	Number of pairs	Mean	S.D.	t-value
Listwise					
	Female	193	73.95	17.92	0.422
	Male	199	73.15	19.31	
Pairwise					
	Female	230	73.72	18.52	0.547
	Male	239	72.72	20.77	
Grand mean					
	Female	242	73.69	18.05	0.547
	Male	255	72.75	20.10	
Cell mean					
	Female	242	73.65	18.35	0.593
	Male	255	72.61	20.43	
Simple regression					
	Female	242	73.98	18.15	0.614
	Male	255	72.92	20.30	

did not demonstrate significantly different posttest level of mastery on criterion-referenced reading tests in the data sets treated with the five missing data techniques. These results support that males and females ended the year at roughly the same level of achievement.

Research Question 7 - Posttest by SES Research Question 7 asked whether student posttest level of mastery scores on criterion-referenced reading tests would differ between high- and low-socioeconomic (SES) students. The research hypothesis addressing this question stated that there would be no significant difference in posttest level of mastery score between high- and low-SES students. Independent t-tests were used to test this hypothesis.

The means, standard deviations and test results for student performance on criterion-referenced reading posttests across the five missing data techniques are displayed in Table 32. The hypothesis that there would be no significant difference in posttest level of mastery scores between high- and low-SES students was not rejected ($t=0.896$; $t=0.994$; $t=0.987$; $t=1.210$; $t=1.132$) in any of the five data sets. In all five data sets, high- and low-SES students did not demonstrate significantly different posttest level of mastery scores on the criterion-referenced reading tests. With the gains by low-SES students allowing for a "closing of the gap," these results support the belief that low- and high-SES students were demonstrating similar levels of achievement at the end of the school year.

Table 32. Analysis of low socioeconomic and middle and high socioeconomic students, criterion-referenced, K-6 reading posttest level of mastery scores, across all grades, by missing data technique

MDT	SES	Number of pairs	Mean	S.D.	t-value
Listwise					
	Middle and high	344	73.86	18.50	0.896
	Low	48	71.29	19.51	
Pairwise					
	Middle and high	404	73.57	19.39	0.994
	Low	65	70.96	21.44	
Grand mean					
	Middle and high	427	73.55	18.86	0.987
	Low	70	71.12	20.66	
Cell mean					
	Middle and high	427	73.54	19.14	1.210
	Low	70	70.51	21.07	
Simple regression					
	Middle and high	427	73.84	19.01	1.132
	Low	70	71.02	20.77	

Research Question 8 - Posttest by level of absence The eighth research question asked whether student posttest level of mastery scores on criterion-referenced reading tests would differ by level of absence. The research hypothesis stated that there would be no significant difference in posttest level of mastery scores between students with high levels of absence and those with normal levels of absence. To test this hypothesis across the five data sets, independent t-tests were used.

Table 33 presents the means, standard deviations, and test results for student posttest level of mastery scores on criterion-referenced reading tests. The hypothesis that there would be no

Table 33. Analysis of high absence and normal attendance student, criterion-referenced, K-6 reading posttest level of mastery scores, across all grades, by missing data technique

MDT	Absence	Number of pairs	Mean	S.D.	t-value
Listwise					
	Normal	306	74.21	17.96	1.234
	High	86	71.17	20.76	
Pairwise					
	Normal	351	73.96	18.79	1.312
	High	118	70.98	22.07	
Grand mean					
	Normal	369	73.92	18.32	1.314
	High	128	71.16	21.19	
Cell mean					
	Normal	369	73.71	18.61	1.087
	High	128	71.39	21.61	
Simple regression					
	Normal	369	74.09	18.45	1.196
	High	128	71.55	21.45	

significant difference in posttest level of mastery scores between students with high and normal levels of absence was not rejected (t=1.234; t=1.312; t=1.314; t=1.087; t=1.196) in any of the five data sets. In all five data sets, high- and normal-absence students did not demonstrate significantly different posttest level of mastery scores on the criterion-referenced reading tests.

Research Question 9 – Posttest by grade The ninth research question asked whether student posttest level of mastery scores differed by grade. The research hypothesis was that there would be no significant difference in posttest level of mastery scores between grades. A one-way analysis of variance (ANOVA) was used to test this hypothesis. Tables 34 through 38 present the means and standard deviations for student posttest performance across the five missing data techniques.

Across all five missing data techniques, the order of the grades was the same. The second grade produced the highest posttest level of mastery score with scores near 90 percent. From high to low, grades one, kindergarten, and third followed with scores in the low to mid 80 percent range. Grades five and six produced the lowest level of mastery scores with performance in the mid 50 percent range.

Table 34. Mean percents and standard deviations in criterion-referenced, K-6 reading posttest level of mastery scores, by grade, from data treated with the listwise deletion missing data technique

Grade	Cases	Mean	S.D.
K	43	84.99	16.00
1	32	86.17	8.78
2	57	90.39	10.20
3	58	81.58	10.54
4	73	72.28	14.39
5	51	54.90	15.93
6	78	57.14	13.10

Table 35. Mean percents and standard deviations in criterion-referenced, K–6 reading posttest level of mastery scores, by grade, from data treated with the pairwise deletion missing data technique

Grade	Cases	Mean	S.D.
K	52	85.80	15.62
1	55	85.81	12.66
2	69	90.14	9.71
3	64	80.94	11.55
4	74	71.69	15.15
5	66	53.22	16.31
6	89	55.46	13.55

Table 36. Mean percents and standard deviations in criterion-referenced, K–6 reading posttest level of mastery scores, by grade, from data treated with the grand-mean substitution missing data technique

Grade	Cases	Mean	S.D.
K	55	85.12	15.45
1	57	85.37	12.65
2	74	88.99	10.30
3	65	80.82	11.50
4	81	71.82	14.48
5	72	54.89	16.56
6	93	56.22	13.74

Table 37. Mean percents and standard deviations in criterion-referenced, K–6 reading posttest level of mastery scores, by grade, from data treated with the cell-mean substitution missing data technique

Grade	Cases	Mean	S.D.
K	55	85.80	15.18
1	57	85.81	12.43
2	74	90.14	9.37
3	65	80.94	11.46
4	81	71.69	14.47
5	72	53.22	15.60
6	93	55.46	13.25

Table 38. Mean percents and standard deviations in criterion-referenced, K–6 reading posttest level of mastery scores, by grade, from data treated with the simple regression missing data technique

Grade	Cases	Mean	S.D.
K	55	85.62	15.21
1	57	85.02	13.14
2	74	89.93	9.57
3	65	80.97	11.46
4	81	72.40	14.80
5	72	54.73	16.53
6	93	56.14	13.73

Table 39 displays the results of the ANOVA of student posttest performance by grade level. With the focus on determining if the choice of missing data technique impacted the results of the ANOVA, the results of the Scheffé multiple range tests for differences between specific pairs of grades are not reported here. Scheffé multiple range tests were performed and the results can be found in Appendix C.

The null hypothesis that there would be no significant difference in posttest level of mastery scores between grades levels was rejected across all five missing data techniques ($F=67.427$, $p=.000$; $F=85.100$, $p=.000$; $F=79.813$, $p=.000$; $F=95.907$, $p=.000$; $F=82.698$, $p=.000$). It can be concluded that, across data treated with five different missing data techniques, posttest level of mastery scores differed significantly by grade level.

Table 39. One-way analysis of variance of criterion-referenced, K-6 reading posttest level of mastery scores, by grade, for all missing data techniques

MDT	Source	df	Sum of squares	Mean squares	F ratio
Listwise					
	Between groups	6	6.948	1.158	67.427***
	Within groups	385	6.612	0.017	
	Total	391	13.559		
Pairwise					
	Between groups	6	9.517	1.586	85.100***
	Within groups	462	8.611	0.019	
	Total	468	18.129		
Grand mean					
	Between groups	6	8.960	1.493	79.813***
	Within groups	490	9.168	0.019	
	Total	496	18.129		
Cell mean					
	Between groups	6	10.113	1.685	95.907***
	Within groups	490	8.611	0.018	
	Total	496	18.724		
Regression					
	Between groups	6	9.270	1.545	82.698***
	Within groups	490	9.155	0.019	
	Total	496	18.425		

***$p<.001$.

Research questions - Math

This section of the chapter reports on the findings connected to the nine research questions in the subject of math.

Research Question 1 - Pretest to posttest The first research question asked whether students' level of mastery scores improved from pretest to posttest. The first hypothesis stated that posttest level of mastery scores would be significantly higher than pretest level of mastery scores. An independent t-test was used to test this hypothesis in data treated with the pairwise deletion missing data technique. Dependent t-tests were used to test the hypothesis in data treated with the other missing data techniques.

Table 40 shows the means, standard deviations, and test results for student performance on criterion-referenced math tests across all five missing data techniques. The null hypothesis that there would be no significant differences in pretest and posttest level of mastery scores in math was rejected across all five missing data techniques ($t=28.679$, $p=.000$; $t=21.969$, $p=.000$; $t=35.584$, $p=.000$; $t=34.878$, $p=.000$; $t=34.514$, $p=.000$). It can be concluded that posttest level of mastery scores on the criterion-referenced math tests are significantly higher than the pretest level of mastery scores on the criterion-referenced math tests across data treated with all five missing data techniques. The students in this study greatly improved their performance in math. These results support that actual learning occurred in these students in the subject of math.

Research Question 2 - Gain by gender The second research question asked whether students' gains in level of mastery on criterion-referenced math tests would differ between males and females. The second research hypothesis stated that there would be no significant difference in gain in math level of mastery between males and females. Independent t-tests were used to test this hypothesis.

Table 40. Analysis of criterion-referenced, mean K–12 math pretest and posttest level of mastery scores, across all grades, by missing data technique

MDT	Test	Number of pairs	Mean	S.D.	t-value
Listwise					
	Pretest	691	47.02	13.53	28.679**
	Posttest	691	64.54	18.34	
Pairwise					
	Pretest	815	46.82	13.60	21.969**
	Posttest	828	64.67	18.94	
Grand mean					
	Pretest	988	46.82	12.35	35.584**
	Posttest	988	64.67	17.34	
Cell mean					
	Pretest	988	46.34	12.56	34.878**
	Posttest	988	63.80	17.63	
Simple regression					
	Pretest	988	46.39	12.67	34.514**
	Posttest	988	63.34	17.99	

**$p < .000$.

Because gains in mastery in data treated with the pairwise deletion missing data technique were generated by subtracting the pretest mean from the posttest mean which did not create a distribution of scores, tests to check for significance of differences could not be performed. This resulted in reporting statistics on only four of the five missing data techniques on research questions addressing gains in level of mastery.

Table 41 shows the mean, standard deviation, and test results for student performance on criterion-referenced math tests across four of the five missing data techniques. The hypothesis that there would be no significant difference in gain in level of mastery scores between males and females was not rejected in any of the data sets ($t=-0.394$; $t=0.250$; $t=0.249$; $t=0.526$). Males and females did not demonstrate significantly different gains in mastery on criterion-referenced math tests in data treated with four of the five missing data techniques. Though none of the differences were significant, females did slightly outperform males in all but the data set treated with listwise deletion. The results support the belief that, in terms of math, males and females were learning equally well during the course of the year.

Table 41. Analysis of male and female gains in criterion-referenced, K-12 math level of mastery scores, across all grades, by missing data technique

MDT	Gender	Number of pairs	Mean	S.D.	t-value
Listwise					
	Female	332	17.22	15.51	-0.394
	Male	359	17.70	16.58	
Pairwise[a]					
	Female	--	18.01	--	
	Male	--	17.70	--	
Grand mean					
	Female	472	17.98	15.28	0.250
	Male	516	17.73	16.21	
Cell mean					
	Female	472	17.59	15.43	0.249
	Male	516	17.34	16.03	
Simple regression					
	Female	472	17.22	15.03	0.526
	Male	516	16.70	15.81	

[a]Gains obtained through direct calculation since no distribution of scores exists.

Research Question 3 - Gain by SES The third research question asked whether students' gain in level of mastery on criterion-referenced math tests would differ between high and low socioeconomic (SES) students. The third research hypothesis stated that there would be no significant difference in gain in math level of mastery between high- and low-SES students. Independent t-tests were used to test this hypothesis.

Table 42 shows the mean, standard deviations, and test results for student performance on criterion-referenced math tests across four of the five missing data techniques. Because gains in mastery in data treated with the pairwise deletion missing data technique were generated by subtracting the pretest mean from the posttest mean which did not create a distribution of scores, a t-test could not be performed.

The hypothesis that there would be no significant difference in gain in level of mastery scores between high- and low-SES students was not rejected in any of the data sets ($t=-0.497$; $t=-1.068$; $t=-1.036$; $t=-1.166$). It can be concluded that high- and low-SES students did not demonstrate significantly different gains in mastery on the criterion-referenced math tests. Though not statistically significant, low-SES students gained more than the middle and high-SES students in all five data sets. These results support that students made similar progress in the area of math during the year regardless of SES.

Research Question 4 - Gain by level of absence The fourth research question asked whether students' gains in level of mastery on criterion-referenced math tests would differ by level of absence. The fourth research hypothesis stated that there would be no significant difference in gain in math level of mastery between students with high levels of absence and normal levels of absence. Independent t-tests were used to test this hypothesis.

Table 42. Analysis of low socioeconomic and middle and high socioeconomic student gains in criterion-referenced, K-12 math level of mastery scores, across all grades, by missing data technique

MDT	SES	Number of pairs	Mean	S.D.	t-value
Listwise					
	Middle and high	628	17.37	15.77	-0.497
	Low	63	18.43	18.88	
Pairwise[a]					
	Middle and high	--	17.60	--	
	Low	--	20.32	--	
Grand mean					
	Middle and high	897	17.64	15.34	-1.068
	Low	91	19.90	19.51	
Cell mean					
	Middle and high	897	17.26	15.30	-1.036
	Low	91	19.45	19.53	
Simple regression					
	Middle and high	897	16.74	15.09	-1.166
	Low	91	19.07	18.47	

[a]Gains obtained through direct calculation since no distribution of scores exists.

*$p < .05$.

Table 43 shows the mean, standard deviation, and test results for student performance on criterion-referenced math tests across four of the five missing data techniques. Because gains in mastery in data treated with the pairwise deletion missing data technique were generated by subtracting the pretest mean from the posttest mean which did not create a distribution of scores, a t-test could not be performed. The hypothesis that there would be no significant difference in gain in math level of mastery scores between students with high levels of absence and normal levels of absence was rejected in every data set ($t=3.00$, $p<.01$; $t=2.754$, $p<.01$; $t=2.801$, $p<.01$; $t=3.003$, $p<.01$). It can be concluded that students with normal levels of absence make significantly higher gains on the criterion-referenced math tests than students with high levels of absence. These results support the position that students who are absent less learn more math over the course of the year.

Table 43. Analysis of high absence and normal attendance student gains in criterion-referenced, K-12 math level of mastery scores, across all grades, by missing data technique

MDT	Absence	Number of pairs	Mean	S.D.	t-value
Listwise					
	Normal	519	18.52	15.90	3.000**
	High	171	14.29	16.25	
Pairwise[a]					
	Normal	--	18.89	--	
	High	--	15.07	--	
Grand mean					
	Normal	681	18.77	15.73	2.754**
	High	301	15.77	15.83	
Cell mean					
	Normal	681	18.42	15.73	2.801**
	High	301	15.37	15.72	
Simple regression					
	Normal	681	17.93	15.51	3.003**
	High	301	14.73	15.21	

[a]Gains obtained through direct calculation since no distribution of scores exists.

**$p<.01$.

Research Question 5 - Gain by grade The fifth research question asked whether students' gains in math level of mastery differed by grade. The fifth research hypothesis stated that there would be no significant difference in gain in level of mastery between grade levels. A one-way analysis of variance (ANOVA) was used to test this hypothesis.

Tables 44–48 show the means and available standard deviations for student performance on the criterion-referenced math tests across the five missing data techniques. Across all five missing data techniques, seventh grade showed the smallest gain, usually around 9 percent. Grades eight and five alternated in second and third lowest position with a range of scores from 9 to 12 percent. The four grades showing the largest gains, from high to low, were second (upper 30 percent range), first (mid 30 percent range), kindergarten (low 20 percent range), and third grade (low 20 percent range). The remaining grades alternated positions across the five data sets between these two ends.

Table 44. Mean percents and standard deviations for gains in criterion-referenced, K–12 math level of mastery scores, by grade, from data treated with the listwise deletion missing data technique

Grade	Cases	Mean	S.D.
K	43	23.57	21.14
1	32	33.19	19.34
2	57	39.95	13.24
3	58	21.08	12.00
4	73	12.53	14.50
5	51	12.22	11.35
6	78	16.41	11.00
7	74	9.05	11.88
8	73	10.51	10.13
9	51	13.33	11.46
10	55	14.44	15.55
11	33	15.70	13.79
12	13	19.08	15.25

Table 45. Mean percents for gains in criterion-referenced, K–12 math level of mastery scores, by grade, from data treated with the pairwise[a] deletion missing data technique

Grade	Cases	Mean	S.D.
K	--	22.94	--
1	--	34.81	--
2	--	39.86	--
3	--	20.76	--
4	--	12.28	--
5	--	9.71	--
6	--	16.14	--
7	--	9.42	--
8	--	10.91	--
9	--	12.81	--
10	--	14.36	--
11	--	16.52	--
12	--	16.85	--

[a]Gains obtained through direct calculation since no distribution of scores exists.

Table 46. Mean percents and standard deviations for gains in criterion-referenced, K–12 math level of mastery scores, by grade, from data treated with the grand-mean substitution missing data technique

Grade	Cases	Mean	S.D.
K	55	22.60	20.34
1	57	34.69	17.85
2	74	38.88	13.76
3	65	21.11	12.57
4	81	12.82	14.79
5	72	10.29	13.31
6	93	15.85	11.94
7	77	9.49	12.07
8	84	10.21	10.12
9	89	14.10	12.20
10	88	15.11	14.23
11	80	17.13	11.53
12	73	18.79	10.31

Table 47. Mean percents and standard deviations for gains in criterion-referenced, K-12 math level of mastery scores, by grade, from data treated with the cell-mean substitution missing data technique

Grade	Cases	Mean	S.D.
K	55	22.94	20.20
1	57	34.81	17.86
2	74	39.86	12.22
3	65	20.76	12.28
4	81	12.28	14.52
5	72	9.71	13.66
6	93	16.14	11.78
7	77	9.42	12.04
8	84	9.91	9.76
9	89	12.81	12.14
10	88	14.36	14.25
11	80	16.52	11.06
12	73	16.85	9.10

Table 48. Mean percents and standard deviations for gains in criterion-referenced, K-12 math level of mastery scores, by grade, from data treated with the simple regression missing data technique

Grade	Cases	Mean	S.D.
K	55	22.98	19.26
1	57	34.00	17.58
2	74	38.08	12.76
3	65	20.70	12.34
4	81	13.28	14.40
5	72	10.23	13.08
6	93	16.21	11.15
7	77	9.43	11.87
8	84	10.63	9.75
9	89	12.92	10.89
10	88	12.32	14.92
11	80	14.18	12.12
12	73	14.71	9.96

Table 49 shows the results of the ANOVA of student performance by grade level. Given the interest in determining if the choice of missing data technique impacted the results of the ANOVA, the results of the Scheffé multiple range tests for differences between specific pairs of grades are not reported here. Scheffé multiple range tests were run and can be found in Appendix D.

Table 49. One-way analysis of variance of gains in criterion-referenced, K–12 math level of mastery scores, by grade, for all missing data techniques[a]

MDT	Source	df	Sum of squares	Mean squares	F ratio
Listwise					
	Between groups	12	5.264	0.439	23.707***
	Within groups	678	12.546	0.019	
	Total	690	17.810		
Grand mean					
	Between groups	12	6.966	0.580	32.207***
	Within groups	975	17.573	0.018	
	Total	987	24.539		
Cell mean					
	Between groups	12	7.594	0.633	36.615***
	Within groups	975	16.851	0.017	
	Total	987	24.445		
Regression					
	Between groups	12	6.894	0.575	33.692***
	Within groups	975	16.626	0.017	
	Total	987	23.520		

[a]Gains obtained through direct calculation since no distribution of scores exists.

***$p < .001$.

The null hypothesis that there would be no significant difference in gain in math level of mastery between grade levels was rejected across all the data sets ($F=23.707$, $p=.000$; $F=32.207$, $p=.000$; $F=36.615$, $p=.000$; $F=33.692$, $p=.000$). It can be concluded that, across data treated with four of the five missing data techniques, gains in level of mastery differed significantly by grade level.

Research Question 6 - Posttest by gender The sixth research question asked whether student posttest level of mastery scores on criterion-referenced math would differ between males and females. The hypothesis stated that there would be no significant difference in posttest math level of mastery scores between males and females. This hypothesis was tested using an independent t-test.

Table 50 displays the mean, standard deviation, and posttest results for student performance on criterion-referenced math tests across the five missing data techniques. The hypothesis that there would be no significant difference in math posttest level of mastery scores between males and females was not rejected in any data set ($t=-0.129$; $t=-0.380$; $t=-0.380$; $t=-0.353$; $t=-0.009$). Males and females did not demonstrate significantly different posttest level of mastery on criterion-referenced math tests in the data sets treated with the five missing data techniques. These results support the belief that males and females ended that year at roughly the same level of math achievement.

Research Question 7 - Posttest by SES Research Question 7 asked whether student posttest level of mastery scores on criterion-referenced math tests would differ between high- and low-socioeconomic (SES) students. The research hypothesis addressing this question stated that there would be no significant difference in math posttest level of mastery score between high- and low-SES students. Independent t-tests were used to test this hypothesis.

Table 50. Analysis of male and female, criterion-referenced, K–12 math posttest level of mastery scores, across all grades, by missing data technique

MDT	Gender	Number of pairs	Mean	S.D.	t-value
Listwise					
	Female	332	64.45	17.33	-0.129
	Male	359	64.63	19.26	
Pairwise					
	Female	397	64.42	17.76	-0.380
	Male	431	64.91	19.98	
Grand mean					
	Female	472	64.46	16.28	-0.380
	Male	516	64.87	18.26	
Cell mean					
	Female	472	63.60	16.61	-0.353
	Male	516	63.99	18.54	
Simple regression					
	Female	472	63.34	16.89	-0.009
	Male	516	63.35	18.95	

The means, standard deviations, and test results for student performance on criterion-referenced math posttests across the five missing data techniques are displayed in Table 51. The hypothesis that there would be no significant difference in posttest level of mastery scores between high- and low-SES students was not rejected (t=0.608; t=0.105; t=0.105; t=-0.272; t=-0.351) in any of the five data sets. In all five data sets, high- and low-SES students did not demonstrate significantly different posttest level of mastery scores on the criterion-referenced math tests. As in the subject of reading, the gains by low-SES students allowed for a "closing of the gap," supporting the belief that low- and high-SES students were demonstrating similar levels of math achievement at the end of the school year.

Table 51. Analysis of low socioeconomic and middle and high socioeconomic students, criterion-referenced, K-12 math posttest level of mastery scores, across all grades, by missing data technique

MDT	SES	Number of pairs	Mean	S.D.	t-value
Listwise					
	Middle and high	628	64.70	17.98	0.608
	Low	63	62.97	21.74	
Pairwise					
	Middle and high	745	64.70	18.49	0.105
	Low	83	64.43	22.73	
Grand mean					
	Middle and high	897	64.70	16.84	0.105
	Low	91	64.44	21.70	
Cell mean					
	Middle and high	897	63.74	17.16	-0.272
	Low	91	64.39	21.90	
Simple regression					
	Middle and high	897	63.26	17.56	-0.351
	Low	91	64.10	21.91	

Research Question 8 - Posttest by level of absence The eighth research question asked whether student posttest level of mastery scores on criterion-referenced math tests would differ by level of absence. The research hypothesis stated that there would be no significant difference in posttest level of mastery scores between students with high levels of absence and those with normal levels of absence. To test this hypothesis across the five data sets, independent t-tests were used.

Table 52 presents the means, standard deviations, and test results for student posttest level of mastery scores on criterion-referenced math tests. The hypothesis that there would be no significant difference in math posttest level of mastery scores between students with high and normal levels of absence was rejected ($t=3.204$, $p<.01$; $t=3.378$, $p<.01$; $t=3.260$, $p<.01$; $t=3.784$, $p<.001$; $t=3.799$, $p<.001$) in each of the five data sets. In all five data sets, students with normal levels of absence possessed significantly higher posttest scores than students with high levels of absence. These results support the position that students who are absent less during the course of the year finish the year with higher levels of math achievement.

Table 52. Analysis of high absence and normal attendance student, criterion-referenced, K-12 math posttest level of mastery scores, across all grades, by missing data technique

MDT	Absence	Number of pairs	Mean	S.D.	t-value
Listwise					
	Normal	519	65.80	18.05	3.204**
	High	171	60.65	18.77	
Pairwise					
	Normal	601	66.02	18.47	3.378**
	High	226	61.05	19.74	
Grand mean					
	Normal	681	65.86	17.36	3.260**
	High	301	61.96	17.17	
Cell mean					
	Normal	681	65.22	17.57	3.784***
	High	301	60.62	17.54	
Simple regression					
	Normal	681	64.78	17.96	3.799***
	High	301	60.07	15.21	

**$p<.01$.

***$p<.001$.

Research Question 9 - Posttest by grade The ninth research question asked whether student math posttest level of mastery scores differed by grade. The research hypothesis was that there would be no significant difference in posttest level of mastery scores between grades. A one-way analysis of variance (ANOVA) was used to test this hypothesis. Tables 53-57 present the means and standard deviations for student posttest performance across the five missing data techniques.

Across all five missing data techniques, the grade with the lowest posttest scores was seventh with scores around the 53 percent range. The next lowest was eighth grade with scores in the mid 50 percent range. Across all data sets, the four grades with the highest posttest scores were, from high to low, second (upper 80 percent range), first grade (mid 80 percent range), third (mid 70 percent range), and kindergarten (upper 60 percent range). The remaining grades

Table 53. Mean percents and standard deviations in criterion-referenced, K-12 math posttest level of mastery scores, by grade, from data treated with the listwise deletion missing data technique

Grade	Cases	Mean	S.D.
K	43	69.61	17.25
1	32	83.82	14.32
2	57	89.15	10.23
3	58	76.59	11.87
4	73	57.46	18.08
5	51	64.75	14.50
6	78	59.44	14.88
7	74	53.08	14.87
8	73	55.01	14.51
9	51	62.98	14.27
10	55	62.65	16.80
11	33	57.58	15.42
12	13	58.77	17.37

Table 54. Mean percents and standard deviations in criterion-referenced, K-12 math posttest level of mastery scores, by grade, from data treated with the pairwise deletion missing data technique

Grade	Cases	Mean	S.D.
K	52	68.91	17.67
1	56	85.08	15.81
2	70	89.19	10.06
3	64	76.13	12.99
4	74	56.99	18.36
5	68	61.05	16.79
6	91	58.66	14.80
7	76	53.42	14.99
8	80	54.46	14.17
9	69	62.38	14.67
10	66	62.61	16.47
11	48	57.33	14.40
12	14	55.43	20.85

Table 55. Mean percents and standard deviations in criterion-referenced, K-12 math posttest level of mastery scores, by grade, from data treated with the grand-mean substitution missing data technique

Grade	Cases	Mean	S.D.
K	55	68.68	17.20
1	57	84.72	15.90
2	74	87.86	11.26
3	65	75.96	12.97
4	81	57.65	17.67
5	72	61.25	16.33
6	93	58.79	14.66
7	77	53.57	14.95
8	84	54.94	14.00
9	89	62.89	12.94
10	88	63.12	14.27
11	80	60.27	11.68
12	73	62.90	9.59

Table 56. Mean percents and standard deviations in criterion-referenced, K-12 math posttest level of mastery scores, by grade, from data treated with the cell-mean substitution missing data technique

Grade	Cases	Mean	S.D.
K	55	68.91	17.17
1	57	85.08	15.67
2	74	89.19	9.78
3	65	76.13	12.89
4	81	56.99	17.54
5	72	61.05	16.31
6	93	58.66	14.64
7	77	53.42	14.89
8	84	54.46	13.83
9	89	62.38	12.90
10	88	62.61	14.24
11	80	57.33	11.11
12	73	55.43	8.86

Table 57. Mean percents and standard deviations in criterion-referenced, K-12 math posttest level of mastery scores, by grade, from data treated with the simple regression missing data technique

Grade	Cases	Mean	S.D.
K	55	69.39	17.31
1	57	84.81	15.80
2	74	88.34	10.57
3	65	78.98	12.95
4	81	57.82	17.82
5	72	60.90	16.48
6	93	58.62	14.71
7	77	53.52	14.92
8	84	54.91	14.02
9	89	61.16	13.43
10	88	59.96	15.60
11	80	55.13	12.22
12	73	55.78	11.43

alternated positions across the five data sets between these two ends with posttest scores ranging from the upper 50s to mid 60s.

Table 58 displays the results of the ANOVA of student posttest performance by grade level. With the focus on determining if the choice of missing data technique impacted the results of the

Table 58. One-way analysis of variance of criterion-referenced, K-12 math posttest level of mastery scores, by grade, for all missing data techniques

MDT	Source	df	Sum of squares	Mean squares	F ratio
Listwise					
	Between groups	12	8.038	0.670	29.922***
	Within groups	678	15.178	0.022	
	Total	690	23.216		
Pairwise					
	Between groups	12	10.568	0.881	37.593***
	Within groups	815	19.093	0.023	
	Total	827	29.661		
Grand mean					
	Between groups	12	9.964	0.830	41.104***
	Within groups	975	19.697	0.020	
	Total	987	29.661		
Cell mean					
	Between groups	12	11.599	0.967	49.359***
	Within groups	975	19.093	0.020	
	Total	987	30.692		
Regression					
	Between groups	12	11.430	0.953	45.280***
	Within groups	975	20.510	0.021	
	Total	987	31.941		

***$p < .001$.

ANOVA, the results of the Scheffé multiple range tests for differences between specific pairs of grades are not reported here. Scheffé multiple range tests were performed and the results can be found in Appendix E.

The null hypothesis that there would be no significant difference in math posttest level of mastery scores between grade levels was rejected across all five missing data techniques ($F=29.922$, $p<.001$; $F=37.593$, $p<.001$; $F=41.104$, $p<.001$; $F=49.359$, $p<.001$; $F=45.280$, $p<.001$). It can be concluded that, across data treated with five different missing data techniques, math posttest level of mastery scores differed significantly by grade level.

Research questions - Language Arts

This section of the chapter reports on the findings connected to the nine research questions in the subject of language arts.

Research Question 1 - Pretest to posttest The first research question asked whether students' level of mastery scores improved from pretest to posttest. The first hypothesis stated that posttest level of mastery scores would be significantly higher than pretest level of mastery scores. An independent t-test was used to test this hypothesis in data treated with the pairwise deletion missing data technique. Dependent t-tests were used to test the hypothesis in data treated with the other missing data techniques.

Table 59 shows the means, standard deviations, and test results for student performance on criterion-referenced language arts tests across all five missing data techniques. The null hypothesis that there would be no significant differences in retest and posttest level of mastery scores in language arts was rejected across all five missing data techniques ($t=15.592$, $p=.000$; $t=11.596$, $p=.000$; $t=20.326$, $p=.000$; $t=19.435$, $p=.000$; $t=21.605$, $p=.000$). It can be concluded that posttest level of mastery scores on the criterion-referenced language tests are significantly higher than the pretest level of mastery scores on the criterion-referenced language arts tests, across data treated with all five missing data techniques. The students in this study greatly improved their performance in language arts. These results support that actual learning occurred in these students in the subject of language arts.

Research Question 2 - Gain by gender The second research question asked whether students' gains in level of mastery on criterion-referenced language arts tests would differ between males and females. The second research hypothesis stated that there would be no

Table 59. Analysis of criterion-referenced, mean K-12 language arts pretest and posttest level of mastery scores, across all grades, by missing data technique

MDT	Test	Number of pairs	Mean	S.D.	t-value
Listwise					
	Pretest	691	54.16	19.50	15.592**
	Posttest	691	65.35	20.32	
Pairwise					
	Pretest	901	51.41	19.47	11.596**
	Posttest	907	62.44	20.95	
Grand mean					
	Pretest	988	51.41	18.59	20.326**
	Posttest	988	62.44	20.07	
Cell mean					
	Pretest	988	51.44	19.14	19.435**
	Posttest	988	62.09	20.72	
Simple regression					
	Pretest	988	51.53	19.03	21.605**
	Posttest	988	62.12	20.49	

**$p < .01$.

significant difference in gain in level of mastery between males and females. Independent t-tests were used to test this hypothesis.

Because gains in mastery in data treated with the pairwise deletion missing data technique were generated by subtracting the pretest mean from the posttest mean which did not create a distribution of scores, tests to check for significance of differences could not be performed. This resulted in reporting statistics on only four of the five missing data techniques on research questions addressing gains in level of mastery.

Table 60 shows the mean, standard deviation, and test results for student performance on criterion-referenced language arts tests across four of the five missing data techniques. The hypothesis that there would be no significant difference in gain in language arts level of mastery scores between males and females was not rejected ($t=0.026$; $t=0.721$; $t=0.502$; $t=0.518$) in any of the data sets. Males and females did not demonstrate significantly different gains in mastery on criterion-referenced language arts tests in data treated with four of the five missing data techniques. Though not significantly, females did slightly outperform males in each of the data sets. The results support that, in terms of language arts, males and females learned equally well over the course of the year.

Research Question 3 – Gain by SES The third research question asked whether students' gain in level of mastery on criterion-referenced language arts tests would differ between high/middle and low socioeconomic (SES) students. The third research hypothesis stated that there would be no significant difference in gain in level of mastery between high/middle- and low-SES students. Independent t-tests were used to test this hypothesis.

Table 61 shows the mean, standard deviations, and test results for student performance on criterion-referenced language arts tests across four of the five missing data techniques. Because gains in mastery in data treated with the pairwise deletion missing data technique were generated

Table 60. Analysis of male and female gains in criterion-referenced, K–12 language arts level of mastery scores, across all grades, by missing data technique

MDT	Gender	Number of pairs	Mean	S.D.	t-value
Listwise					
	Female	332	11.18	15.37	0.026
	Male	359	11.15	17.86	
Pairwise[a]					
	Female	--	11.49	--	
	Male	--	10.66	--	
Grand mean					
	Female	472	11.44	15.85	0.721
	Male	516	10.66	18.01	
Cell mean					
	Female	472	10.95	16.09	0.502
	Male	516	10.40	18.24	
Simple regression					
	Female	472	10.86	14.45	0.518
	Male	516	10.35	16.25	

[a]Gains obtained through direct calculation since no distribution of scores exists.

by subtracting the pretest mean from the posttest mean which did not create a distribution of scores, a t-test could not be performed.

The hypothesis that there would be no significant difference in gain in language arts level of mastery scores between high/middle- and low-SES students was not rejected ($t=-1.167$; $t=-1.251$; $t=-1.424$; $t=-1.122$) in any of the data sets. High/middle- and low-SES students did not demonstrate significantly different gains in mastery on criterion-referenced language arts tests in data treated with four of the five missing data techniques. These results support that low-SES students made the same progress in the area of reading as did other students.

Table 61. Analysis of low socioeconomic and middle and high socioeconomic student gains in criterion-referenced, K-12 language arts level of mastery scores, across all grades, by missing data technique

MDT	SES	Number of pairs	Mean	S.D.	t-value
Listwise					
	Middle and high	628	10.93	16.85	-1.167
	Low	63	16.50	15.08	
Pairwise[a]					
	Middle and high	--	10.79	--	
	Low	--	13.28	--	
Grand mean					
	Middle and high	897	10.82	17.14	-1.251
	Low	91	13.16	16.16	
Cell mean					
	Middle and high	897	10.41	17.24	-1.424
	Low	91	13.11	17.13	
Simple regression					
	Middle and high	897	10.42	15.59	-1.122
	Low	91	12.32	13.54	

[a]Gains obtained through direct calculation since no distribution of scores exists.

*$p < .05$.

Research Question 4 - Gain by level of absence The fourth research question asked whether students' gains in level of mastery on criterion-referenced language arts tests would differ by level of absence. The fourth research hypothesis stated that there would be no significant difference in gain in level of mastery between students with high levels of absence and normal levels of absence. Independent t-tests were used to test this hypothesis.

Table 62 shows the mean, standard deviation, and test results for student performance on criterion-referenced language arts tests across four of the five missing data techniques. Because gains in mastery in data treated with the pairwise deletion missing data technique were generated by subtracting the pretest mean from the posttest mean which did not create a distribution of scores, a t-test could not be performed. The hypothesis that there would be no significant difference in gain in language arts level of mastery scores between students with high levels of absence and normal levels of absence was not rejected ($t=1.638$; $t=1.079$; $t=1.454$; $t=1.213$) in

Table 62. Analysis of high absence and normal attendance student gains in criterion-referenced, K–12 language arts level of mastery scores, across all grades, by missing data technique

MDT	Absence	Number of pairs	Mean	S.D.	t-value
Listwise					
	Normal	519	11.78	16.55	1.638
	High	71	9.37	17.09	
Pairwise[a]					
	Normal	--	11.35	--	
	High	--	9.83	--	
Grand mean					
	Normal	681	11.39	17.02	1.079
	High	301	10.12	17.22	
Cell mean					
	Normal	681	11.24	17.33	1.454
	High	301	9.50	17.13	
Simple regression					
	Normal	681	10.96	15.62	1.213
	High	301	9.67	15.01	

[a]Gains obtained through direct calculation since no distribution of scores exists.

any of the data sets. Students with high levels of absence and normal levels of absence did not demonstrate significantly different gains in mastery on criterion-referenced language arts tests in data treated with four of the five missing data techniques. Though not significant, students with normal levels of absence showed larger gains than students with high levels of absence across all data sets.

Research Question 5 - Gain by grade The fifth research question asked whether students' gains in language arts level of mastery differed by grade. The fifth research hypothesis stated that there would be no significant difference in gain in level of mastery between grade levels. A one-way analysis of variance (ANOVA) was used to test this hypothesis.

Tables 63-67 show the means and available standard deviations for student performance on the criterion-referenced language arts tests across the five missing data techniques.

Table 63. Mean percents and standard deviations for gains in criterion-referenced, K-12 language arts level of mastery scores, by grade, from data treated with the listwise deletion missing data technique

Grade	Cases	Mean	S.D.
K	43	45.45	17.63
1	32	32.55	15.21
2	57	15.88	11.46
3	58	17.24	11.98
4	73	6.82	13.19
5	51	5.74	7.61
6	78	7.16	12.63
7	74	5.25	8.06
8	73	8.35	10.40
9	51	6.35	10.28
10	55	-1.31	10.05
11	33	2.54	9.11
12	13	10.00	10.83

Table 64. Mean percents for gains in criterion-referenced, K–12 language arts level of mastery scores, by grade, from data treated with the pairwise[a] deletion missing data technique

Grade	Cases	Mean	S.D.
K	--	44.25	--
1	--	36.18	--
2	--	15.33	--
3	--	16.85	--
4	--	6.94	--
5	--	4.13	--
6	--	6.49	--
7	--	5.24	--
8	--	8.27	--
9	--	4.47	--
10	--	-1.88	--
11	--	2.70	--
12	--	10.91	--

[a]Gains obtained through direct calculation since no distribution of scores exists.

Table 65. Mean percents and standard deviations for gains in criterion-referenced, K–12 language arts level of mastery scores, by grade, from data treated with the grand-mean substitution missing data technique

Grade	Cases	Mean	S.D.
K	55	41.39	18.55
1	57	31.83	15.24
2	74	16.73	15.04
3	65	17.41	12.66
4	81	6.87	13.05
5	72	6.10	16.03
6	93	6.95	13.18
7	77	5.31	9.06
8	84	8.43	10.44
9	89	4.73	10.90
10	88	-1.47	13.89
11	80	8.28	16.39
12	73	9.96	12.45

Table 66. Mean percents and standard deviations for gains in criterion-referenced, K-12 language arts level of mastery scores, by grade, from data treated with the cell-mean substitution missing data technique

Grade	Cases	Mean	S.D.
K	55	44.25	17.96
1	57	36.18	16.64
2	74	15.33	11.50
3	65	16.85	12.72
4	81	6.94	13.07
5	72	4.13	16.97
6	93	6.49	13.43
7	77	5.24	9.05
8	84	8.27	10.43
9	89	4.47	10.74
10	88	-1.88	11.35
11	80	2.70	10.13
12	73	10.90	12.26

Table 67. Mean percents and standard deviations for gains in criterion-referenced, K-12 language arts level of mastery scores, by grade, from data treated with the simple regression missing data technique

Grade	Cases	Mean	S.D.
K	55	39.99	19.23
1	57	26.39	14.93
2	74	15.40	10.96
3	65	17.09	11.87
4	81	7.36	12.84
5	72	6.26	15.73
6	93	7.45	11.83
7	77	5.25	8.86
8	84	8.49	9.92
9	89	5.14	10.42
10	88	-2.15	10.95
11	80	5.88	11.62
12	73	10.21	11.33

Across all five missing data techniques, tenth grade showed the lowest change in score, and the only negative one, with a one to two percent drop. The six grades with the largest gains, in order from highest to lowest across all five data sets, were: kindergarten (40 to 45 percent range), first (mid 20 to mid 30 percent range), third (mid teens range), second (mid teens range), twelfth grade (10 percent range), and eighth grade (8 percent range). The remaining grades alternated positions across the five data sets between these two ends with gains ranging from 2 to 8 percent.

Table 68 shows the results of the ANOVA of student performance by grade level. Given the interest in determining if the choice of missing data technique impacted the results of the ANOVA, the results of the Scheffé multiple range tests for differences between specific pairs of grades are not reported here. Scheffé multiple range tests were run and can be found in Appendix F.

The null hypothesis that there would be no significant difference in gain in language arts level of mastery between grade levels was rejected across the data sets ($F=47.703$, $p<.001$; $F=47.627$, $p<.001$; $F=68.455$, $p<.001$; $F=47.588$, $p<.001$). It can be concluded that, across data treated with four of the five missing data techniques, gains in language arts level of mastery differed significantly by grade level.

Research Question 6 - Posttest by gender The sixth research question asked whether student posttest level of mastery scores on criterion-referenced language arts would differ between males and females. The hypothesis stated that there would be no significant difference in posttest language arts level of mastery scores between males and females. This hypothesis was tested using an independent t-test.

Table 68. One-way analysis of variance of gains in criterion-referenced, K–12 language arts level of mastery scores, by grade, for all missing data techniques[a]

MDT	Source	df	Sum of squares	Mean squares	F ratio
Listwise					
	Between groups	12	8.810	0.734	47.703***
	Within groups	678	10.434	0.015	
	Total	690	19.244		
Grand mean					
	Between groups	12	10.617	0.885	47.627***
	Within groups	975	18.112	0.019	
	Total	987	28.729		
Cell mean					
	Between groups	12	13.414	1.118	68.455***
	Within groups	975	15.921	0.016	
	Total	987	29.334		
Regression					
	Between groups	12	8.660	0.722	47.588***
	Within groups	975	14.786	0.015	
	Total	987	23.447		

[a]Pairwise deletion did not produce a distribution of scores and was excluded.

***$p < .001$.

Table 69 displays the mean, standard deviation, and posttest results for student performance on criterion-referenced language arts tests across the five missing data techniques. The hypothesis that there would be no significant difference in language arts posttest level of mastery scores between males and females was rejected in data treated with pairwise deletion, grand-mean substitution, cell-mean substitution, and simple regression missing data techniques ($t=2.34$, $p<.05$; $t=2.347$, $p<.05$; $t=1.974$, $p<.05$; $t=2.097$, $p<.05$). The hypothesis that there would be no significant difference in language arts posttest level of mastery between males and females was not rejected (t=1.761) in data treated with the listwise deletion missing data technique. This may be due to the reduced number of subjects, and resulting loss of power, connected with the use of listwise deletion.

In the majority of data sets, females demonstrated a significantly higher posttest level of mastery on criterion-referenced language arts tests than males. Females did demonstrate higher language arts posttest scores across all five data sets. These results support that females ended the year ahead of males in terms of achievement in language arts.

Table 69. Analysis of male and female, criterion-referenced, K-12 language arts posttest level of mastery scores, across all grades, by missing data technique

MDT	Gender	Number of pairs	Mean	S.D.	t-value
Listwise					
	Female	332	66.76	19.81	1.761
	Male	359	64.05	20.72	
Pairwise					
	Female	428	64.16	20.27	2.344*
	Male	479	60.90	21.45	
Grand mean					
	Female	472	64.01	19.30	2.347*
	Male	516	61.01	20.67	
Cell mean					
	Female	472	63.45	20.14	1.974*
	Male	516	60.85	21.17	
Simple regression					
	Female	472	63.55	19.76	2.097*
	Male	516	60.82	21.07	

*$p < .05$.

Research Question 7 - Posttest by SES Research Question 7 asked whether student posttest level of mastery scores on criterion-referenced language arts tests would differ between high/middle- and low-socioeconomic (SES) students. The research hypothesis addressing this question stated that there would be no significant difference in language arts posttest level of mastery scores between high/middle- and low-SES students. Independent t-tests were used to test this hypothesis.

The means, standard deviations, and test results for student performance on criterion-referenced language arts posttests across the five missing data techniques are displayed in Table 70. The hypothesis that there would be no significant difference in language arts posttest level of mastery scores between high/middle- and low-SES students was rejected in data treated with pairwise deletion, grand-mean substitution, cell-mean substitution, and simple regression missing data techniques ($t=-2.031$, $p<.05$; $t=-2.025$, $p<.05$; $t=-2.835$, $p<.01$; $t=-2.371$, $p<.05$). The hypothesis that there would be no significant difference in language arts posttest level of mastery between high/middle- and low-SES students was not rejected ($t=-0.669$) in data treated with the listwise deletion missing data technique. This may be due to the reduced number of subjects, and resulting loss of power, connected with the use of listwise deletion.

In the majority of data sets, low-SES demonstrated a significantly higher posttest level of mastery on criterion-referenced language arts tests than high and middle-SES students. Low-SES students did demonstrate higher language arts posttest scores across all five data sets. These results support that low-SES students ended the year ahead of high and middle-SES students in terms of achievement in language arts.

Research Question 8 - Posttest by level of absence The eighth research question asked whether student posttest level of mastery scores on criterion-referenced language arts tests would differ by level of absence. The research hypothesis stated that there would be no significant

Table 70. Analysis of low socioeconomic and middle and high socioeconomic students, criterion-referenced, K-12 language arts posttest level of mastery scores, across all grades, by missing data technique

MDT	SES	Number of pairs	Mean	S.D.	t-value
Listwise					
	Middle and high	628	65.22	20.72	-0.669
	Low	63	66.67	15.84	
Pairwise					
	Middle and high	825	62.05	21.20	-2.031*
	Low	82	66.35	17.97	
Grand mean					
	Middle and high	897	62.09	20.33	-2.025*
	Low	91	65.97	17.09	
Cell mean					
	Middle and high	897	61.58	20.93	-2.835**
	Low	91	67.21	17.74	
Simple regression					
	Middle and high	897	61.69	20.70	-2.371*
	Low	91	66.39	17.73	

*$p<.05$.

**$p<.01$.

difference in posttest level of mastery scores between students with high levels of absence and those with normal levels of absence. To test this hypothesis across the five data sets, independent t-tests were used.

Table 71 presents the means, standard deviations, and test results for student posttest level of mastery scores on criterion-referenced language arts tests. The hypothesis that there would be no significant difference in math posttest level of mastery scores between students with high and

Table 71. Analysis of high absence and normal attendance student, criterion-referenced, K-12 language arts posttest level of mastery scores, across all grades, by missing data technique

MDT	Absence	Number of pairs	Mean	S.D.	t-value
Listwise					
	Normal	519	67.68	19.42	5.360***
	High	71	58.26	21.41	
Pairwise					
	Normal	642	65.12	20.41	6.141***
	High	264	55.90	20.90	
Grand mean					
	Normal	681	64.97	19.82	6.043***
	High	301	56.70	19.68	
Cell mean					
	Normal	681	64.97	20.14	6.452***
	High	301	55.90	20.68	
Simple regression					
	Normal	681	64.78	20.18	6.195***
	High	301	56.14	20.11	

***$p < .001$.

normal levels of absence was rejected (t=5.360, p<.001; t=6.141, p<.001; t=6.043, p<.001; t=6.452, p<.001; t=6.195, p<.001) in each of the five data sets. In all five data sets, students with normal levels of absence possessed significantly higher posttest scores than students with high levels of absence. These results support the position that students who are absent less during the course of the year finish the year with higher levels of achievement in language arts.

Research Question 9 - Posttest by grade The ninth research question asked whether student language arts posttest level of mastery scores differed by grade. The research hypothesis was that there would be no significant difference in posttest level of mastery scores between grades. A one-way analysis of variance (ANOVA) was used to test this hypothesis. Tables 72-76 present the means and standard deviations for student posttest performance across the five missing data techniques.

Across all five missing data techniques, eleventh and tenth grade alternated the lowest two scores, with posttest scores falling in the mid 30 to low 40 percent range. The third through seventh lowest grades were the same, from lowest to highest, twelfth (mid to upper 40 percent range), ninth (low to mid 50 percent range), seventh (low 60 percent range), eighth (low 60 percent range), and sixth (low 60 percent range). The top grade in terms of language arts posttest scores was second (low 90 percent range) followed by third grade (low 80 percent range). The remaining grades alternated positions across the five data sets between these two ends with posttest scores falling in the upper 60 to the mid 70 percent range.

Table 72. Mean percents and standard deviations in criterion-referenced, K-12 language arts posttest level of mastery scores, by grade, from data treated with the listwise deletion missing data technique

Grade	Cases	Mean	S.D.
K	43	72.56	23.15
1	32	71.22	18.07
2	57	92.02	8.18
3	58	82.76	10.34
4	73	70.33	17.92
5	51	74.51	15.92
6	78	63.95	13.41
7	74	61.29	13.15
8	73	62.37	11.67
9	51	54.51	13.17
10	55	40.18	13.35
11	33	34.07	10.18
12	13	45.38	8.02

Table 73. Mean percents and standard deviations in criterion-referenced, K–12 language arts posttest level of mastery scores, by grade, from data treated with the pairwise deletion missing data technique

Grade	Cases	Mean	S.D.
K	53	72.99	22.42
1	42	71.43	16.45
2	69	91.74	7.80
3	64	82.38	11.09
4	74	69.84	18.29
5	66	71.75	15.67
6	87	62.94	14.12
7	76	61.30	13.02
8	80	62.35	11.35
9	85	52.35	13.34
10	82	39.24	13.57
11	60	35.50	10.82
12	69	48.41	13.95

Table 74. Mean percents and standard deviations in criterion-referenced, K–12 language arts posttest level of mastery scores, by grade, from data treated with the grand-mean substitution missing data technique

Grade	Cases	Mean	S.D.
K	55	72.61	22.10
1	57	69.06	17.11
2	74	89.76	10.56
3	65	82.07	11.28
4	81	69.20	17.60
5	72	71.02	15.20
6	93	62.91	13.65
7	77	61.32	12.93
8	84	62.36	11.08
9	89	52.81	13.20
10	88	40.83	14.36
11	80	42.23	15.01
12	73	49.17	13.93

Table 75. Mean percents and standard deviations in criterion-referenced, K-12 language arts posttest level of mastery scores, by grade, from data treated with the cell-mean substitution missing data technique

Grade	Cases	Mean	S.D.
K	55	72.99	22.01
1	57	71.43	16.64
2	74	91.74	7.53
3	65	82.38	11.00
4	87	69.84	17.47
5	72	71.75	14.99
6	93	62.94	13.65
7	77	61.30	12.93
8	84	62.35	11.08
9	89	52.35	13.03
10	88	39.24	13.10
11	80	35.50	9.35
12	73	48.41	13.55

Table 76. Mean percents and standard deviations in criterion-referenced, K-12 language arts posttest level of mastery scores, by grade, from data treated with the simple regression missing data technique

Grade	Cases	Mean	S.D.
K	55	72.64	22.23
1	57	65.13	20.28
2	74	91.02	8.64
3	65	82.35	11.01
4	81	70.26	17.79
5	72	71.19	15.50
6	93	63.00	13.83
7	77	61.20	12.96
8	84	62.44	11.25
9	89	52.81	13.28
10	88	40.33	14.01
11	80	39.03	11.59
12	73	48.97	13.77

Table 77 displays the results of the ANOVA of student posttest performance by grade level. With the focus on determining if the choice of missing data technique impacted the results of the ANOVA, the results of the Scheffé multiple range tests for differences between specific pairs of grades are not reported here. Scheffé multiple range tests were performed and the results can be found in Appendix G.

Table 77. One-way analysis of variance of criterion-referenced, K-12 language arts posttest level of mastery scores, by grade, for all missing data techniques

MDT	Source	df	Sum of squares	Mean squares	F ratio
Listwise					
	Between groups	12	14.787	1.232	61.021***
	Within groups	678	13.691	0.020	
	Total	690	28.478		
Pairwise					
	Between groups	12	21.379	1.782	86.577***
	Within groups	894	18.397	0.021	
	Total	906	39.775		
Grand mean					
	Between groups	12	19.247	1.604	76.178***
	Within groups	975	20.529	0.021	
	Total	987	39.776		
Cell mean					
	Between groups	12	23.964	1.997	105.841***
	Within groups	975	18.397	0.019	
	Total	987	42.361		
Regression					
	Between groups	12	21.119	1.760	84.534***
	Within groups	975	20.299	0.021	
	Total	987	41.419		

***$p < .001$.

The null hypothesis that there would be no significant difference in language arts posttest level of mastery scores between grades levels was rejected across all five missing data techniques ($F=61.021$, $p<.001$; $F=86.577$, $p<.001$; $F=76.178$, $p<.001$; $F=105.841$, $p<.001$; $F=84.534$, $p<.001$). It can be concluded that, across data treated with five different missing data techniques, language arts posttest level of mastery scores differed significantly by grade level.

Part 2

The second part of this study consisted of conducting a simulation to investigate the performance of the five missing data techniques used in the first part of the study. The goal was to study the feasibility of using the simulation to help districts determine which missing data technique would be appropriate to use for any given set of data.

The complete set

The simulation began with a complete set of data, with absolutely no missing data. The set used was the one generated using the listwise deletion method in part one of the study. The completeness of the data set allowed its use as a "comparison" set. Tables 78-80 display the means and standard deviations on pretest, posttest, and gain scores across all three subjects. These are the same as those reported for the data treated with listwise deletion in part one of the study.

Table 81 holds the intercorrelations between the criterion-referenced pretest and posttest level of mastery scores, across all grades and subjects, from data in the complete data set. It, too, matches the results provided from data treated with the listwise deletion missing data technique from part one of the study.

Table 78. Mean percents and standard deviations for criterion-referenced pretest level of mastery scores, across all grades, for all subjects in the complete data set

Subject	Cases	Mean	S.D.
Reading	392	60.92	17.62
Math	691	47.02	13.53
Language Arts	691	54.16	19.50

Table 79. Mean percents and standard deviations for criterion-referenced posttest level of mastery scores, across all grades, for all subjects in the complete data set

Subject	Cases	Mean	S.D.
Reading	392	73.54	18.62
Math	691	64.54	18.34
Language Arts	691	65.35	20.32

Table 80. Mean percents and standard deviations for gains in criterion-referenced level of mastery scores, across all grades, for all subjects in the complete data set

Subject	Cases	Mean	S.D.
Reading	392	12.62	15.40
Math	691	17.47	16.07
Language Arts	691	11.16	16.70

Table 81. Intercorrelations among criterion-referenced pretest and posttest level of mastery scores, across all grades, for all subjects from data in the complete data set

Variable	Reading pretest	Reading posttest	Math pretest	Math posttest	LA pretest	LA posttest
Reading pretest	--	.640	.534	.550	.477	.641
Reading posttest	--	--	.381	.656	.200	.610
Math pretest	--	--	--	.527	.429	.421
Math posttest	--	--	--	--	.364	.607
LA pretest	--	--	--	--	--	.648
LA posttest	--	--	--	--	--	--

Note. $n=392$ for correlations involving reading tests scores, $n=691$ for all others; $p<.001$ for all cases.

The deviation measures

The data sets to which the missing data techniques were applied were generated by randomly removing data from the complete set. The data removal pattern was set to match, within each subject, the proportion of missing data by grade level and type of test. A total of ten "proportionally equivalent" data sets were created from the initial complete data set.

Performance of the missing data techniques was judged on the magnitude of deviation measures. The smaller the deviation measure, the better the missing data technique performed. Two deviation measures were used to track missing data technique performance. The mean deviation measures was calculated by summing the differences between the actual and estimated, or imputed, values and dividing by the number of differences. The mean absolute deviation measures were calculated by summing the absolute value of the differences between the actual and estimated, or imputed, values and dividing by the number of differences.

The measure deviation measure, which balanced negatives and positives, was used to track the bias of the missing data techniques. The mean absolute deviation measure, which indicated the positive "distances," was used to track how "close" estimates were to the actual values. For both measures, scores were averaged across the ten proportional data sets.

Performance in estimating means The mean deviation scores for pretest and posttest level of mastery means is displayed in Table 82. In general, the missing data techniques produced mean deviation scores of small magnitude. Across all subjects and missing data techniques, there were only four instances when the mean deviation score had a magnitude greater than 0.4 percent. These included the math posttest score in data treated with the simple regression missing

Table 82. Mean deviation scores for criterion-referenced, pretest and posttest level of mastery score means, for all subjects, across all proportionally equivalent data sets, by missing data technique

Subject	Test	Listwise	Pairwise	Grand mean	Cell mean	Regression
Reading						
	Pre	0.116	-0.079	-0.079	0.012	0.020
	Post	0.106	0.097	0.067	0.012	0.062
Math						
	Pre	0.107	0.210	0.210	0.099	-0.002
	Post	-0.685	0.276	0.296	0.021	-0.578
Language Arts						
	Pre	1.103	-0.008	-0.008	0.117	0.137
	Post	1.074	0.268	0.268	0.000	0.363
Mean		0.304	0.127	0.126	0.044	0.000[a]

[a]Unrounded value equals 0.0003.

data technique (−0.578 percent) and the listwise deletion missing data technique (−0.685 percent). Data treated with the listwise deletion method produced the largest mean deviation scores on language arts pretest (1.103 percent) and posttest (1.074 percent) level of mastery scores. The results presented in Table 82 can also be found in graph form in Appendix H.

In terms of the mean deviation score, the simple regression missing data technique performs the best across all subjects and tests in estimating test means (0.000 percent). The cell-mean substitution method ranked second in performance on the mean deviation scores, followed by grand mean and pairwise deletion missing data techniques. As determined by the mean deviation scores, the listwise deletion missing data technique performed the worst in estimating test means, averaging a difference of approximately 0.3 percent.

Table 83 displays the values of the mean absolute deviation scores for pretest and posttest level of mastery scores. In terms of the mean absolute deviation scores, the cell-mean substitution missing data technique performed the best across all subjects and test in estimating test means (0.181 percent). The grand-mean substitution method ranked second in performance (0.208 percent) on the mean absolute deviation scores, followed closely by the pairwise deletion (0.209 percent) and simple regression (0.240 percent) missing data techniques. As with the mean deviation scores, the listwise deletion missing data technique performed the worst on mean absolute deviation scores in estimating test means (0.657 percent). The listwise deletion missing data technique produced the largest mean absolute deviation scores across all test scores. The results presented in Table 83 can also be found in graph form in Appendix H.

Performance in estimating standard deviations The second statistic that deviation measures were calculated for were standard deviations for the pretest and posttest level of mastery scores across all subjects. The overall mean deviation measures indicated that listwise deletion, pairwise deletion, grand-mean substitution, and cell-mean substitution all produced estimates that

Table 83. Mean absolute deviation scores for criterion-referenced, pretest and posttest level of mastery score means, for all subjects, across all proportionally equivalent data sets, by missing data technique

Subject	Test	Listwise	Pairwise	Grand mean	Cell mean	Regression
Reading						
	Pre	0.344	0.171	0.171	0.136	0.084
	Post	0.416	0.139	0.109	0.102	0.114
Math						
	Pre	0.319	0.210	0.210	0.151	0.116
	Post	0.685	0.278	0.298	0.275	0.578
Language Arts						
	Pre	1.103	0.190	0.190	0.301	0.187
	Post	1.074	0.268	0.268	0.120	0.363
Mean		0.657	0.209	0.208	0.181	0.240

were lower than the standard deviations in the complete data set. The mean deviation measures from the simple regression missing data technique constantly overestimated the standard deviations.

In terms of the mean deviation measure, the pairwise deletion missing data technique did the best job (-0.035) of estimating standard deviations of test score distributions. Next in terms of performance on the mean deviation measure were the simple regression (0.224), listwise deletion (-0.389), and cell-mean substitution (-0.543) missing data techniques. The grand-mean substitution method was the worst at estimating standard deviations with a mean deviation of -0.758. Table 84 presents the mean deviation scores for pretest and posttest level of mastery score standard deviations. The results presented in Table 84 can also be found in graph form in Appendix H.

Table 84. Mean deviation scores for criterion-referenced, pretest and posttest level of mastery score standard deviations, for all subjects, across all proportionally equivalent data sets, by missing data technique

Subject	Test	Listwise	Pairwise	Grand mean	Cell mean	Regression
Reading						
	Pre	-0.400	-0.076	-0.578	-0.643	0.034
	Post	-0.411	-0.036	-0.541	-0.274	0.114
Math						
	Pre	0.034	0.101	-0.787	-0.686	0.116
	Post	0.029	0.108	-0.867	-0.645	0.578
Language Arts						
	Pre	-0.609	-0.150	-0.981	-0.746	0.137
	Post	-0.976	-0.154	-0.793	-0.265	0.363
Mean		-0.389	-0.035	-0.758	-0.543	0.224

Similar findings exist among the mean absolute deviation scores. By mean absolute deviation measure, the pairwise deletion missing data technique did the best job in estimating test score standard deviations. The simple regression (0.224), listwise deletion (-0.487), and cell-mean substitution (0.543) missing data techniques were next in estimating standard deviations. The grand-mean substitution method did the worst job in estimating standard deviation with a mean absolute deviation measure of 0.830. The mean absolute deviation scores for the pretest and posttest level of mastery scores are presented in Table 85. The results presented in Table 85 can also be found in graph form in Appendix H.

Performance in estimating correlations Deviation measures were generated for correlations in order to investigate the impact the various missing data techniques had on the relationships between test scores. The mean deviation scores for correlations between test scores

Table 85. Mean absolute deviation scores for criterion-referenced, pretest and posttest level of mastery score standard deviations, for all subjects, across all proportionally equivalent data sets, by missing data technique

Subject	Test	Listwise	Pairwise	Grand mean	Cell mean	Regression
Reading						
	Pre	0.344	0.171	0.171	0.136	0.084
	Post	0.416	0.139	0.109	0.102	0.114
Math						
	Pre	0.319	0.210	0.210	0.151	0.116
	Post	0.685	0.278	0.298	0.275	0.578
Language Arts						
	Pre	1.103	0.190	0.190	0.301	0.187
	Post	1.074	0.268	0.268	0.120	0.363
MEAN		0.487	0.124	0.830	0.543	0.224

are displayed in Table 86. The pairwise deletion, grand-mean substitution, and cell-mean substitution missing data techniques tended to underestimate the relationship between test scores. The simple regression missing data technique usually overestimated the relationship between test scores. The results presented in Table 86 can also be found in graph form in Appendix H.

Based on mean deviation scores, the listwise deletion missing data technique was best (.0034) at estimating the relationship between test scores. The simple regression (.0134), cell-mean substitution (−.0223), and grand-mean substitution (−.0325) missing data techniques were the middle performers. The pairwise deletion missing data technique, with a mean deviation score of −.0406, did the worst job of estimating the relationships between tests scores.

Mean absolute deviation measures for correlations between pretest and posttest level of mastery scores are given in Table 87. With a mean absolute deviation score of .0608, the

Table 86. Mean deviation scores for intercorrelations among criterion-referenced, pretest and posttest level of mastery scores, for all subjects, across all proportionally equivalent data sets, by missing data technique

Pair	Listwise	Pairwise	Grand mean	Cell mean	Regression
RePr & RePs	.0019	.0655	.0415	.0122	-.0027
RePr & MaPr	.0047	-.0452	-.0434	-.0417	.0077
RePr & MaPs	-.0136	-.0565	-.0368	-.0193	.0067
RePr & LAPr	-.0050	-.0589	-.0567	-.0451	-.0055
RePr & LAPs	.0000	-.0648	-.0515	-.0197	.0095
RePs & MaPr	.0098	-.0361	-.0251	-.0230	.0175
RePs & MaPs	-.0032	-.0666	-.0332	-.0155	.0145
RePs & LAPr	-.0013	-.0181	-.0143	-.0393	.0214
RePs & LAPs	.0078	-.0626	-.0344	-.0162	-.0080
MaPr & MaPs	.0167	-.0452	-.0240	-.0172	.0178
MaPr & LAPr	.0112	-.0410	-.0464	-.0345	.0067
MaPr & LAPs	.0185	-.0350	-.0361	-.0182	.0248
MaPs & LAPr	.0017	-.0368	-.0359	-.0212	.0388
MaPs & LAPs	.0176	-.0402	-.0419	-.0191	.0433
LAPr & LAPs	-.0151	-.0671	-.0495	-.0170	.0091
Mean	.0034	-.0406	-.0325	-.0223	.0134

Note. Re=Reading, Ma=Math, LA=Language Arts, Pr=Pretest, Ps=Posttest.

Table 87. Mean absolute deviation scores for intercorrelations among criterion-referenced, pretest and posttest level of mastery scores, for all subjects, across all proportionally equivalent data sets, by missing data technique

Pair	Listwise	Pairwise	Grand mean	Cell mean	Regression
RePr & RePs	.0165	.0713	.0415	.0134	.0175
RePr & MaPr	.0239	.0702	.0434	.0417	.0179
RePr & MaPs	.0176	.0613	.0368	.0193	.0083
RePr & LAPr	.0200	.0605	.0567	.0451	.0193
RePr & LAPs	.0160	.0730	.0515	.0197	.0195
RePs & MaPr	.0264	.0433	.0251	.0230	.0175
RePs & MaPs	.0114	.0736	.0332	.0161	.0145
RePs & LAPr	.0245	.0257	.0143	.0393	.0220
RePs & LAPs	.0198	.0674	.0344	.0162	.0136
MaPr & MaPs	.0191	.0634	.0240	.0186	.0202
MaPr & LAPr	.0154	.0548	.0464	.0345	.0173
MaPr & LAPs	.0207	.0518	.0361	.0200	.0248
MaPs & LAPr	.0057	.0436	.0359	.0212	.0388
MaPs & LAPs	.0176	.0812	.0419	.0191	.0433
LAPr & LAPs	.0151	.0705	.0495	.0170	.0175
Mean	.0180	.0608	.0380	.0243	.0208

Note. Re=Reading, Ma=Math, LA=Language Arts, Pr=Pretest, Ps=Posttest.

pairwise deletion method did the worst job of estimating the relationships between test scores. The listwise deletion missing data technique did the best job of estimating the relationships between test scores with a mean absolute deviation score of .0180. The simple regression (.0208), cell-mean substitution (.0243), and grand-mean substitution (.0380) missing data techniques fell between the other two techniques. The results presented in Table 87 can also be found in graph form in Appendix H.

CHAPTER V. SUMMARY, CONCLUSIONS, LIMITATIONS, DISCUSSION, AND RECOMMENDATIONS

This investigation consisted of two parts. The first was to examine how performance of K-12 students on criterion-referenced reading, math, and language arts tests varied in relation to the use of five different missing data techniques. The second part of this study was to investigate a possible method to assist districts in making a missing data technique selection. This was accomplished through the running of a simulation built on the work of Kalton (1983).

Summary

Student performance data were collected on students from Lincoln County School District No. 1 for the 1995-1996 school year. Pretest and posttest scores on criterion-referenced tests in reading (K-6), math (K-12), and language arts (K-12) were used as indicators of student achievement. Gain scores were calculated by subtracting pretest scores from posttest scores. Because test lengths varied and to allow comparison among grades, scores were converted to a "percent correct" indicator referred to as a "level of mastery" score.

The study focused on the performance of five different missing data techniques. These included: listwise deletion, pairwise deletion, grand-mean substitution, cell-mean substitution, and simple regression.

Part 1

The first part of the study tracked the impact of missing data selection on student achievement by monitoring performance on nine research questions in the subjects of reading, math, and language arts. This summary presents the findings of those tested hypotheses. Results for Hypotheses 2, 3, 4, and 5 do not include test results for data treated with the pairwise deletion missing data technique. Because gains were calculated directly from overall posttest and pretest averages, no distribution of gain scores were available on which to apply statistical procedures.

Hypothesis 1 This hypothesis stated that posttest level of mastery scores would be significantly higher than pretest level of mastery scores in all subjects. The null hypothesis that there would be no significant difference posttest and pretest level of mastery scores was tested using a dependent t-test. In every subject, across data sets treated with five different missing data techniques, the null hypothesis was rejected. It was concluded that the posttest level of mastery scores were significantly higher than the pretest level of mastery scores in reading, math, and language arts. This indicates that the choice of missing data technique did not impact the results connected to this hypothesis.

Hypothesis 2 This hypothesis stated that there would be no significant difference in gain in level of mastery score between males and females in all subjects. This hypothesis was tested using an independent t-test. Across all data sets and subjects, the null hypothesis was not rejected. It was concluded that males and females gained the same amount in all subjects. This indicates that the choice of missing data technique did not affect the results connected to this hypothesis.

Hypothesis 3 This hypothesis stated that there would be no significant difference in gain in level of mastery score between high/middle- and low-SES students in any subject. This hypothesis was tested using an independent t-test. In math and language arts, across all data sets, the null hypothesis was not rejected. It was concluded that SES did not impact the gain students made in math and language arts. In reading, in the data treated with the simple regression, the null hypothesis was not rejected. In the remaining data sets the null hypothesis was rejected.

These indicators show that, for reading, the choice of missing data technique did impact the results of this hypothesis.

Hypothesis 4 This hypothesis stated that there would be no significant difference in gain in level of master score between students with high levels of absence and normal levels of absence in any subject. This hypothesis was tested using an independent t-test. In reading and language arts, across all data sets, the null hypothesis was not rejected. It was concluded that, for reading and language arts, gains were the same for all students, regardless of absence level. In math across all data sets, the null hypothesis was rejected. It was concluded that, in math, students with average levels of absence gained more than students with high level of absence. These results indicate that the choice of missing data technique did not affect the results connected to this hypothesis.

Hypothesis 5 This hypothesis stated that there would be no significant difference in gain in level of mastery score between grade levels in any subject. This hypothesis was tested using a one-way analysis of variance (ANOVA). In every subject, across all data sets, the null hypothesis was rejected. It was concluded that the gains in level of mastery score did vary by grade level for each subject. This indicates that the choice of missing data technique did not impact the results connected to this hypothesis.

Hypothesis 6 This hypothesis stated that there would be no significant difference in posttest level of mastery score between males and females in any subject. This hypothesis was tested using an independent t-test. In reading and math, across all data sets, the null hypothesis was not rejected. It was concluded that, for reading and math, gains were the same for males and females. In language arts, in data treated with pairwise deletion, grand-mean substitution, cell-mean substitution, and simple regression, the hypothesis was rejected. In the data set treated with listwise deletion, the hypothesis was not rejected. These results indicate that the choice of missing data technique did affect the results connected to this hypothesis.

Hypothesis 7 This hypothesis stated that there would be no significant difference in posttest level of mastery score between high/middle- and low-SES students in any subject. This hypothesis was tested using an independent t-test. In reading and math, across all data sets, the null hypothesis was not rejected. It was concluded that, for reading and math, gains were the same for high/middle- and low-SES students. In language arts, in data treated with pairwise deletion, grand-mean substitution, cell-mean substitution, and simple regression, the hypothesis was rejected. In the data set treated with listwise deletion, the hypothesis was not rejected. These results indicate that the choice of missing data technique did affect the results connected to this hypothesis.

Hypothesis 8 This hypothesis stated that there would be no significant difference in posttest level of mastery score between students with high levels of absence and normal levels of absence in any subject. This hypothesis was tested using an independent t-test. In math and language arts, across all data sets, the null hypothesis was rejected. It was concluded that, in math and language arts, posttest level of mastery scores were significantly higher for students with normal levels of absence than for students with high levels of absence. In reading across all data sets, the null hypothesis was not rejected. It was concluded that, in reading, posttest level of mastery scores were not affected by level of absence. These results indicate that the choice of missing data technique did not affect the results connected to this hypothesis.

Hypothesis 9 This hypothesis stated that there would be no significant difference in posttest level of mastery score between grade levels on any subject. This hypothesis was tested using a one-way analysis of variance (ANOVA). In every subject, across all data sets, the null hypothesis was rejected. It was concluded that posttest level of mastery scores did vary by grade level for each subject. This indicates that the choice of missing data technique did not impact the results connected to this hypothesis.

Part 2

The second part of the study centered around the running of a simulation. First, a complete set of data was needed to use for comparison purposes. In this study, the data set created using the listwise deletion method across all three subjects was used. Second, the proportions of missing achievement data in the original data set from Lincoln County were recorded, by subject, by the total amount of missing test data as well as type of test data missing. These proportions were used to create ten proportionally equivalent data sets. Finally, the ten proportionally equivalent data sets were created by randomly removing data from the complete set until the proportion of missing data matched those of the original data set, a procedure which was repeated ten times.

Two types of deviation scores were used as indicators of missing data technique performance. The mean deviation score was calculated by summing the differences between all actual and estimated values and dividing by the number of differences. It was used as an indicator of bias in estimated scores. The mean absolute deviation score was calculated by summing the absolute value of the differences between all actual and estimated values and dividing by the number of differences. Since this value was always nonnegative, it was used as a measure of how "close" estimated or imputed values were to actual values. The smaller the magnitude of deviation measure, the better the missing data technique was found to "perform."

Test means Though most techniques showed some over- and underestimation, the five missing data techniques showed positive overall mean deviation scores thus overestimating test means. Using the mean deviation scores, the simple regression missing data technique performed the best in estimating test means while the listwise deletion method performed the worst. In terms of the mean absolute deviation score, the cell-mean substitution was the "closest" at estimating test means, while the listwise deletion method was once again found to perform the worst.

Standard deviations With the exception of the simple regression missing data technique which consistently overestimated standard deviations, the remaining missing data techniques underestimated standard deviations according to the mean deviation scores. The pairwise deletion method performed best, in terms of mean deviation score, in estimating standard deviations, while the grand-mean substitution missing data technique performed the worst. The results of the mean absolute deviation scores showed similar results, with pairwise deletion being the "closest" and grand-mean substitution being the worst in estimating standard deviations.

Correlations According to the mean deviation scores, the pairwise deletion, grand-mean substitution, and cell-mean substitution missing data techniques tended to underestimate correlations while the simple regression missing data technique usually overestimated correlations. Based on mean deviation scores, the listwise deletion missing data technique performed the best in estimating correlations, while the pairwise deletion missing data technique did the worst job. The results of the mean absolute deviation scores showed similar results, with listwise deletion being the "closest" and pairwise deletion being the worst in estimating correlations.

Conclusions

This study addresses two major research questions. The first research question asked, *Would the use of different missing data techniques affect the results found with the data from Lincoln County School District No.1 in Diamondville, Wyoming?* Given that the choice of missing data technique led to differing results in three of the nine hypotheses, the answer to the first research question is yes.

The second research question asked, *How well do different data techniques perform?* In keeping with the findings in the literature, no one missing data technique performed well on both deviation measures in estimating means, standard deviations, and correlations. Given the performance on the deviation measures, the answer to the second research question is, it varies. Table 88 presents the ranking, from best to worst, of the five missing data techniques across the deviation measures.

Table 88. Ranked performance of the five missing data techniques across mean deviation and mean absolute deviation scores for means, standard deviations, and correlations

	Means		Standard deviations		Correlations	
Ranking	Mean deviation	Mean absolute deviation	Mean deviation	Mean absolute deviation	Mean deviation	Mean absolute deviation
Best	Regress	Cell	Pairwise	Pairwise	Listwise	Listwise
	Cell	Grand	Regress	Regress	Regress	Regress
	Grand	Pairwise	Listwise	Listwise	Cell	Cell
	Pairwise	Regress	Cell	Cell	Grand	Grand
Worst	Listwise	Listwise	Grand	Grand	Pairwise	Pairwise

Note. Listwise = listwise deletion, Pairwise = pairwise deletion, Grand = grand-mean substitution, Cell = cell-mean substitution, Regress = simple regression.

If the interest was in estimating test means, the results of this study support the use of cell-mean substitution or simple regression. Should the interest shift to estimating standard deviations, pairwise deletion is the recommended choice. Listwise deletion is the missing data technique that should be utilized if there is interest in estimating correlations. If a district was limited in its ability to apply multiple missing data techniques, the one technique that performs best across the six deviation measures in this study is simple regression.

Limitations

A number of limitations were created by the design of this study. They included:

1. The data were taken from a previous academic year, which prevented alleviating missing data problems.
2. The study focused on only five of dozens of possible techniques. Though the techniques were chosen using input from the literature, ease of use, and variety in approach, they do not represent the entire spectrum of missing data techniques.
3. The second part of the study focused on reviewing a method for investigating missing data technique performance that districts might be able to use. This resulted in the use of individual computer-based software. This prevented the investigation of missing data techniques requiring large computer capacity or statistical modeling.
4. The study utilized data available from the 1995-1996 Lincoln County data set and was limited to available demographic data. Additional variables of interest in terms of reviewing performance may be of interest in the future.

Discussion

Velotta (1995) stated that missing data were an "undeniable characteristic" of educational research and this study was no exception. Use of listwise deletion across all three subjects resulted in a 30 percent drop in student data, from 988 to 691.

Part 1

In reviewing the results from Part 1 of the study, each subject is discussed in terms of the presence or lack of differences that were found across the data sets treated with the five missing data techniques.

Reading If a district were to review the findings for reading across the data sets treated with the five missing data techniques, it would find that the results would be very similar. There would be significant improvement from pretest to posttest on the criterion-referenced reading tests and that gains did not vary by gender or level of absence. With the exception of the data set treated with the simple regression missing data technique, gains would not have been shown to differ by SES. Gains did vary by grade level. Performance on posttest scores were found not to differ by gender, SES, or level of absence. There was a significant difference in posttest performance by grade level.

Math In reviewing the results in math across the data sets treated with the five missing data techniques, there would be no differences at all. Significant gains from pretest and posttest were found. Gains did not differ by gender or SES. Gains did differ by level of absence and grade level. Posttest performance saw similar results with no differences by gender or SES, but with differences by level of absence and grade level. The similarity in results found in reading and particularly in math match the findings of Velotta (1995), Ward and Clark (1991), and Witta and Kaiser (1991) where few if any differences were found in missing data technique performance.

Language Arts It is in reviewing the results for language arts that a district would find some differences. As with the other subjects, there were significant gains from pretest to posttest. These gains did not differ by gender, SES, or level of absence. There were significant differences by grade level. Posttest performance did vary by level of absence. With the exception of the data set treated with the listwise deletion missing data technique, posttest performance was found to differ by SES and level of absence. All data sets showed differences by grade level. These results are in line with the findings of Roth, Campion, and Jones (1996) and Ward and Clark (1991), where different missing techniques were found to impact results.

If a district used listwise deletion in the treatment of missing data, its review of posttest performance would have identified differences by grade level and level of absence, but not have identified differences by the demographic variables of gender and SES. This could be a potential problem in today's world of accountability where educators are supposed to work toward identifying and eliminating differences. In fact, it appears that the technique that is easiest to use and most often used has the potential to actually miss differences.

In the results for language arts, data treated with listwise deletion twice failed to identify differences that were found in data treated with the other missing data techniques. The smaller number of students may have decreased the power of the t-tests run to investigate differences by gender and SES. This illustrates the condition where reduced sample size reduces the power, or ability to identify differences, mentioned by Raymond (1987), Roth (1994), and Roth, Campion and Jones (1996). The magnitude of statistics generated from data treated by listwise deletion tended to be among the smallest, especially in the F statistic generated in the one-way analysis of variance tests.

Part 2

The second part of the study dealt with the performance of a simulation and the calculation of deviation measures. The mean deviation measure was used to track bias, or the overestimation or underestimation, associated with each missing data technique. The mean absolute deviation was used to monitor how "close" the missing data techniques came to estimating the actual values. In both cases, the smaller the magnitude of the deviation measure, the better the deviation measure was found to perform. Deviation measures were calculated for subject test means, standard deviations, and correlations.

Listwise deletion In terms of bias, listwise deletion showed varied results. Listwise deletion performed the worst in terms of estimating test means, overestimating five of the six test means by an average of approximately 0.3 percent. Standard deviations were underestimated four of the six times with an average deviation of -0.389, which ranked it in the middle of the pack in terms of performance. Little if any bias was shown in the estimation of correlations where the overall mean deviation was .0034, which placed listwise deletion as the best estimator of correlation. Beaton (1997), Little and Rubin (1990), Little and Schenker (1995), Raymond (1987), and Roth (1994) all mention the possibility of bias developing using listwise deletion.

In terms of accuracy or closeness of the estimates, listwise deletion showed similar results. It performed the worst in estimating means with a mean absolute deviation score of 0.657. It ranked third in terms of estimating standard deviations with a mean absolute deviation score of 0.487. Listwise deletion did the best job of the missing data techniques in estimating correlations with a mean absolute deviation score of .0180. The lack of matching test means parallels the observation from Velotta (1995) that listwise deletion failed to produce results that matched actual results.

Pairwise deletion Of the five missing techniques in the study, pairwise deletion was the only one for which some of the test statistics could not be calculated. The lack of a distribution of scores for gains led to the inability to calculate t-tests and an analysis of variance test for differences in gain scores. This difficulty matches the observations of Little and Rubin (1987), Raymond (1987), and Velotta (1995) that the use of pairwise deletion can lead to difficulty in calculations and applying statistical procedures.

Pairwise deletion showed a wide range of performance in terms of bias. It was the best technique in estimating standard deviations, with a mean deviation score of -0.035. It ranked fourth in terms of estimating means, overestimating five out of six, with a mean deviation score of 0.127. Pairwise deletion was the worst technique at estimating correlations, underestimating 14 of 15 for a mean deviation score of -.0406.

Pairwise deletion did come the closet to estimating standard deviations, with a mean absolute deviation score of 0.124. It ranked third in closeness to means with a mean absolute deviation score of 0.209. As in terms of bias, pairwise deletion did the worst job of estimating correlations with a mean absolute deviation score of .0608.

Grand-mean substitution Grand-mean substitution faired poorly in terms of bias. It ranked worst in estimating standard deviations, underestimating all six, with a mean deviation score of -0.758. This tendency to underestimate variation is commented on by Little (1987), Little and Rubin (1990), Raymond (1987), Roth (1994), Roth, Campion and Jones (1996), Velotta (1995), and Witta and Kaiser (1991). Underestimating 14 of 15 correlations, pairwise deletion ranked fourth in terms of estimating correlation with a mean deviation score of -.0325. Little and Schenker (1995), Roth (1994), Velotta (1995), Ward and Clark (1991), and Witta and Kaiser (1991) all commented on this tendency. Grand-mean substitution ranked in the middle of the pack in terms of estimating means with a mean deviation score of 0.126.

In terms of closeness, grand-mean substitution did rank second in estimating means with a mean absolute deviation score of 0.208. Performance on correlations matched its ranking in bias with a mean absolute deviation score of .0381. As in bias, grand-mean substitution ranked last in estimating standard deviations, with a mean absolute deviation score of 0.830.

Cell-mean substitution Cell-mean substitution has been said to be an improvement on grand-mean substitution (Little & Rubin, 1990; Little & Schenker, 1995), with the issues of underestimation of variation and correlation present but to a lesser degree (Little & Rubin, 1987; Little & Rubin, 1990; Velotta, 1995). Such was the case in this study.

Cell-mean substitution always ranked one position better than grand-mean substitution. In terms of bias, cell-mean substitution ranked second in estimating means (mean deviation score of 0.044), fourth in terms of standard deviations (mean deviation score of -0.543), and third in estimating correlations (mean deviation score of -0.022).

Similar findings occurred in terms of closeness with cell-mean substitution being best at estimating means with a mean absolute deviation score of 0.181. It ranked third in estimating correlations (mean absolute deviation score of 0.024) and fourth in estimating standard deviations (mean absolute deviation score of 0.543).

Simple regression In terms of bias, simple regression was the most consistent performer. Simple regression did the best job of estimating means with almost no bias given its mean deviation score of 0.000. Being the only missing data technique to always overestimate standard deviation, simple regression ranked second in estimating standard deviations with a mean deviation score of 0.224. It also ranked second in estimating correlations with a mean deviation score of 0.013. The tendency for simple regression to overestimate variation in this study is contrary to findings in the literature, where Beaton (1997) and Roth (1994) both reported a tendency for underestimation.

Simple regression ranked second in terms of closeness in estimating both standard deviations and correlations, with mean absolute deviation scores of 0.224 and 0.021, respectively. Simple regression ranked fourth in closeness in estimating means, with a mean absolute deviation score of 0.240.

Comparisons Velotta (1995) found that, overall, grand-mean substitution and cell-mean substitution produced the lowest standard deviations. This is true in this study, with grand-mean substitution producing the largest underestimation followed by the underestimation produced with cell-mean substitution. Simple regression and cell-mean substitution did do the best job of estimating means, which also matches the observations of Velotta.

In this study, listwise deletion produced the worst estimates of means with the imputation methods producing the better estimates. This is contrary to the findings of Thran and Gillis (1992) in which imputation methods, with the exception of grand-mean substitution, produced higher estimates than listwise deletion. Listwise did show itself to be the best estimator of correlations with pairwise deletion being the worst. This is just the opposite of findings reported by Kim and Curry (1977), where deviations from established correlations showed pairwise deletion superior to listwise deletion.

Overall performance The results show that there is no one missing data technique that outperforms the rest on all of the deviation measures. A total of four different techniques are ranked best according to one or more of the deviation measures. Of note, however, is the overall performance of simple regression. In five of the six cases, simple regression ranks first or second. This strong performance supports the inclusion of simple regression in any set of missing data techniques applied to a set of data.

Of interest is the performance of the most popular missing data technique, listwise deletion. Though it does the best job of estimating correlations, its performance drops to third in the estimation of standard deviations, and last in estimating test means. These results would suggest that listwise deletion should not be the sole missing data technique applied to a set of data, particularly if test means are of primary interest.

Except where noted, the results of this study match the findings in the literature. This consistency may actually help districts in interpreting results from multiple missing data techniques. Findings that are contrary within the missing data techniques applied to a set of data as well as those that are contrary to what is known from the literature should give districts pointers to areas that further investigation is warranted.

Recommendations for Practice

There are several recommendations that come out of this research. The first is that districts, schools, and teachers all need to be aware of the impact that missing data can have on results and work hard to avoid it occurring in the first place. There is no substitution for complete accurate data. Given today's emphasis on results and accountability, it seems only pertinent that efforts should be made to educate researchers and alleviate this problem as much as possible.

Second, the application of missing data techniques should be done in the same time frame as the analysis of results. This will allow the district to follow up quickly if necessary to address missing data issues, i.e., a class missed a test, weather closed some of the schools in the district, papers were misplaced, etc. In addition, it will allow for a quick feel for how confident decision makers should be in the results.

Third, given the apparent tendency to utilize listwise deletion and its demonstrated poor performance in estimating means, academic programs designed to prepared school administrators and researchers should incorporate discussions and work on the issue of missing data. This should help future information providers and decision makers in education with a better feel for how missing data can impact results. It is necessary to move the decision on how to treat missing data from a default to a conscious one.

Fourth, districts reviewing student performance data should incorporate the use of multiple missing data techniques. Though it has been found that there is no one missing data technique that uniformly performs the best, multiple perspectives or angles on an issue will raise awareness and will hopefully add clarity to results. Agreement among multiple missing data techniques can add confidence to interpretation results, while differences can raise questions surrounding possible causes. Given the performance of the techniques in this study, this researcher would recommend including listwise deletion, pairwise deletion, cell-mean substitution, and simple regression. Grand-mean substitution was excluded because of its tendency to show greater underestimation of standard deviations and overall poorer performance than cell-mean substitution without bringing any identified benefit.

Recommendations for Future Research

This study demonstrated the ability to pull from fields outside education to help address issues educators face today. Although there are outstanding researchers within education, building on the work from the fields of statistics and survey research should allow for increased learning. Though apparently not widely addressed, missing data is an issue that can impact student achievement results (with concomitant high-stakes consequences) and should continue to be investigated.

Specifically, it is recommended that research be conducted into finding simulations or other tools that can help districts and buildings investigate missing data technique performance. This may require the generation of a series of approaches, each geared toward investigating one or more aspects of student achievement (means, standard deviations, correlations, etc.).

Another area for suggested research would be an investigation of how the use of multiple missing data techniques to a single set of data impacts results and interpretation. Given the current lack of a best technique, perhaps there are combinations that do not conflict with one another that could meet the needs of educators. Research in this area would help support districts looking for accurate results from being accused of "tweaking" the results to get the best of all worlds.

Research should be conducted into how the various types and number of variables often collected by districts interplay with the application of various missing data techniques. This study focuses only on the characteristics of gender, SES, level of absence, and grade level in terms of posttest scores and gain scores. It would be of interest to know if more continuous variables such as income level, or increasing the number of variables under investigation would change the impact different missing data techniques have on results.

Another area for future research would be to investigate what kinds of criteria perform best in judging the performance of missing data techniques. This study focused on individual computer-based software, making replication of the study and use of the techniques more available to school districts. As more sophisticated statistical software is developed for use, guidance as to adequate criteria for performance would help districts and state education agencies make decisions about the quality of results generated.

APPENDIX A. DISTRICT LETTER AND HUMAN SUBJECTS APPROVAL

LINCOLN COUNTY
SCHOOL DISTRICT NO. 1
Education Specialists
11 Adaville Dr • P.O. Box 335 • (307) 877-9095
DIAMONDVILLE, Wyoming 83116 • FAX # (307) 877-9638

April 16, 1998

Richard Manatt, Director
School Improvement Model (SIM)
Project Office
N225 Lagomarcino Hall
Iowa State University
Ames, IA 50011

Dear Dr. Manatt,

This is to authorize you and your SIM team (specifically, Mr. David Putz, graduate student) to analyze Lincoln County School District No. 1 student and teacher data centered on kindergarten through 12th grade criterion-referenced tests in reading, language arts, and math. We understand that you are investigating techniques for handling missing data that will enable our district to examine advantages and disadvantages of five different missing data techniques.

The district requests a report of all analyses and a description of the findings from the study. Further, it is understood that Mr. Putz will prepare these results as part of his Ph.D. dissertation.

All tests will remain the property of this district. After the acceptance of Mr. Putz's dissertation and the district's reception of the agreed upon reports, SIM is to destroy all reports, data sets, and drafts not returned to the District for deposit in order to assure confidentiality. In-district use of SIM's reports will follow the established human subjects in research regulations of the Board of Education and the State of Wyoming.

If subsequent analysis of any data is requested by the district, it is agreed that an amended Human Subjects in Research request will be made by the School Improvement Model Projects office to both the District and to Iowa State University.

Sincerely,

Terry Ebert, Superintendent

CREATING A PROMISING FUTURE THROUGH EDUCATION

Information for Review of Research Involving Human Subjects

Iowa State University

(Please type and use the attached instructions for completing this form)

1. Title of Project The effects of missing data technique selection on student performance data results.

2. I agree to provide the proper surveillance of this project to insure that the rights and welfare of the human subjects are protected. I will report any adverse reactions to the committee. Additions to or changes in research procedures after the project has been approved will be submitted to the committee for review. I agree to request renewal of approval for any project continuing more than one year.

David J. Putz — 11/30/98 — David J Putz

Typed name of principal investigator — Date — Signature of principal investigator

Educational Leadership & Policy Studies — N243 Lagomarcino

Department — Campus address

Campus: 294-9995 (Dr. Manatt's office)

Phone number to report results

3. Signatures of other investigators — Date — Relationship to principal investigator

Richard P. Manatt — 12-3-98 — Major Professor

4. Principal investigator(s) (check all that apply)
☐ Faculty ☐ Staff ☒ Graduate student ☐ Undergraduate student

5. Project (check all that apply)
☐ Research ☒ Thesis or dissertation ☐ Class project ☐ Independent Study (490, 590, Honors project)

6. Number of subjects (complete all that apply)
____ # adults, non-students ____ # ISU students 600 # minors under 14 ____ other (explain)
400 # minors 14 - 17

7. Brief description of proposed research involving human subjects: (See instructions, item 7. Use an additional page if needed.)

The purpose of this study is to examine the relationship between missing data technique selection and student performance data results. The data comes from district-created criterion-referenced tests in reading, language arts, and mathematics. The source of the data is Lincoln County School District #1 in Diamondville, Wyoming, for the 1995-1996 school year. This data will be used to create simulated sets with which comparisons will be made. Statistical procedures will be used to examine differences in student performance based on missing data technique selection. No students or teachers will be directly contacted.

(Please do not send research, thesis, or dissertation proposals.)

8. Informed Consent:
☐ Signed informed consent will be obtained. (Attach a copy of your form.)
☐ Modified informed consent will be obtained. (See instructions, item 8.)
☒ Not applicable to this project.

9. Confidentiality of Data: Describe below the methods you will use to ensure the confidentiality of data obtained. (See instructions, item 9.)

At no time will individuals be identifiable from the results or reports generated from this study. Numbers are used soley for organization of the information. The interest of this study is the comparison of results based on missing data technique selection, not individuals. The data for this study has been used in a previous study (see 'Process/product research for K-12 schools," submitted by David J. Wilkerson on or around 6/20/96), and will be used with permission from school district administration.

10. What risks or discomfort will be part of the study? Will subjects in the research be placed at risk or incur discomfort? Describe any risks to the subjects and precautions that will be taken to minimize them. (The concept of risk goes beyond physical risk and includes risks to subjects' dignity and self-respect as well as psychological or emotional risk. See instructions, item 10.)

No risks or discomforts are foreseen in this research project.

11. **CHECK ALL** of the following that apply to your research:

- ☐ A. Medical clearance necessary before subjects can participate
- ☐ B. Administration of substances (foods, drugs, etc.) to subjects
- ☐ C. Physical exercise or conditioning for subjects
- ☐ D. Samples (blood, tissue, etc.) from subjects
- ☐ E. Administration of infectious agents or recombinant DNA
- ☐ F. Deception of subjects
- ☒ G. Subjects under 14 years of age and/or ☒ Subjects 14 - 17 years of age
- ☐ H. Subjects in institutions (nursing homes, prisons, etc.)
- ☒ I. Research must be approved by another institution or agency (Attach letters of approval)

If you checked any of the items in 11, please complete the following in the space below (include any attachments):

Items A–E Describe the procedures and note the proposed safety precautions.

Items D–E The principal investigator should send a copy of this form to Environmental Health and Safety, 118 Agronomy Lab for review.

Item F Describe how subjects will be deceived; justify the deception; indicate the debriefing procedure, including the timing and information to be presented to subjects.

Item G For subjects under the age of 14, indicate how informed consent will be obtained from parents or legally authorized representatives as well as from subjects.

Items H–I Specify the agency or institution that must approve the project. If subjects in any outside agency or institution are involved, approval must be obtained prior to beginning the research, and the letter of approval should be filed.

Items G & I: Since the data are property of Lincoln County School District #1 and come from a past school year, consent to use the data was sought from district level administration. A copy of a letter from the Superintendent of Schools is attached to this form.

Last name of Principal Investigator Putz

Checklist for Attachments and Time Schedule

The following are attached (please check):

12. ☐ Letter or written statement to subjects indicating clearly:
    a) the purpose of the research
    b) the use of any identifier codes (names, #s), how they will be used, and when they will be removed (see item 17)
    c) an estimate of time needed for participation in the research
    d) if applicable, the location of the research activity
    e) how you will ensure confidentiality
    f) in a longitudinal study, when and how you will contact subjects later
    g) that participation is voluntary; nonparticipation will not affect evaluations of the subject

13. ☐ Signed consent form (if applicable)

14. ☒ Letter of approval for research from cooperating organizations or institutions (if applicable)

15. ☐ Data-gathering instruments

16. Anticipated dates for contact with subjects:

First contact	Last contact
Fall 1998	Fall 1999
Month/Day/Year	Month/Day/Year

17. If applicable: anticipated date that identifiers will be removed from completed survey instruments and/or audio or visual tapes will be erased:

Fall 1999
Month/Day/Year

18. Signature of Departmental Executive Officer — [signature] Date 12/7/98 Department or Administrative Unit ELPS

19. Decision of the University Human Subjects Review Committee:

☐ Project approved ☐ Project not approved ☐ No action required

Patricia M. Keith
Name of Committee Chairperson — Date — Signature of Committee Chairperson

GC 06/96

APPENDIX B. SCHEFFÉ MULTIPLE RANGE RESULTS FOR GAINS IN K-6 READING LEVEL OF MASTERY SCORES, BY GRADE

Table B1. Scheffé multiple range results for gain in criterion-referenced, K-6 reading level of mastery scores, by grade, from data treated with the listwise deletion missing data technique

	Grade	5	4	6	3	2	K	1
Mean	Grade							
-2.21	5	–						
7.64	4	*	–					
10.85	6	*		–				
14.09	3	*			–			
16.25	2	*	*			–		
24.44	K	*	*	*	*		–	
26.94	1	*	*	*	*	*		–

*$p < .05$.

Table B2. Scheffé multiple range results for gain in criterion referenced, K–6 reading level of mastery scores, by grade, from data treated with the grand-mean substitution missing data technique

	Grade	5	4	6	3	2	K	1
Mean	Grade							
-2.47	5	–						
7.84	4	*	–					
8.93	6	*		–				
14.64	3	*			–			
16.36	2	*	*			–		
23.47	K	*	*	*			–	
25.25	1	*	*	*	*	*		–

*$p < .05$.

Table B3. Scheffé multiple range results for gain in criterion-referenced, K-6 reading level of mastery scores, by grade, from data treated with the cell-mean substitution missing data technique

	Grade	5	4	6	3	2	K	1
Mean	Grade							
-3.71	5	–						
7.54	4	*	–					
9.28	6	*		–				
14.21	3	*			–			
15.65	2	*	*			–		
24.09	K	*	*	*	*	*	–	
25.09	1	*	*	*	*	*		–

*$p < .05$.

Table B4. Scheffé multiple range results for gain in criterion-referenced, K-6 reading level of mastery scores, by grade, from data treated with the simple regression missing data technique

	Grade	5	4	6	3	2	K	1
Mean	Grade							
-2.25	5	--						
8.26	4	*	--					
9.43	6	*		--				
13.82	3	*			--			
15.46	2	*				--		
23.36	K	*	*	*	*		--	
24.51	1	*	*	*	*	*		--

*$p < .05$.

APPENDIX C. SCHEFFÉ MULTIPLE RANGE RESULTS FOR K-6 READING POSTTEST LEVEL OF MASTERY SCORES, BY GRADE

Table C1. Scheffé multiple range results for criterion-references, K–6 reading posttest level of mastery scores, by grade, from data treated with the listwise deletion missing data technique

Mean	Grade	5	6	4	3	K	1	2
54.90	5	–						
57.14	6		–					
72.28	4	*	*	–				
81.58	3	*	*	*	–			
84.99	K	*	*	*		–		
86.17	1	*	*	*			–	
90.39	2	*	*	*	*			–

*$p < .05$.

Table C2. Scheffé multiple range results for criterion-referenced, K–6 reading posttest level of mastery scores, by grade, from data treated with the pairwise deletion missing data technique

	Grade	5	6	4	3	K	1	2
Mean	Grade							
53.22	5	–						
55.46	6		–					
71.69	4	*	*	–				
80.94	3	*	*	*	–			
85.80	K	*	*	*		–		
85.81	1	*	*	*			–	
90.14	2	*	*	*	*			–

*$p < .05$.

Table C3. Scheffé multiple range results for criterion-referenced, K–6 reading posttest level of mastery scores, by grade, from data treated with the grand-mean substitution missing data technique

	Grade	5	6	4	3	K	1	2
Mean	Grade							
54.89	5	--						
56.22	6		--					
71.82	4	*	*	--				
80.82	3	*	*	*	--			
85.12	K	*	*	*		--		
85.37	1	*	*	*			--	
88.99	2	*	*	*				--

*$p < .05$.

Table C4. Scheffé multiple range results for criterion-referenced, K–6 reading posttest level of mastery scores, by grade, from data treated with the cell-mean substitution missing data technique

	Grade	5	6	4	3	K	1	2
Mean	Grade							
53.22	5	–						
55.46	6		–					
71.69	4	*	*	–				
80.94	3	*	*	*	–			
85.80	K	*	*	*		–		
85.81	1	*	*	*			–	
90.14	2	*	*	*	*			–

*p<.05.

Table C5. Scheffé multiple range results for criterion-referenced, K–6 reading posttest level of mastery scores, by grade, from data treated with the simple regression missing data technique

	Grade	5	6	4	3	1	K	2
Mean	Grade							
54.73	5	–						
56.14	6		–					
72.40	4	*	*	–				
80.97	3	*	*	*	–			
85.02	1	*	*	*		–		
85.62	K	*	*	*			–	
89.93	2	*	*	*	*			–

*$p < .05$.

APPENDIX D. SCHEFFÉ MULTIPLE RANGE RESULTS FOR GAINS IN K-12 MATH LEVEL OF MASTERY SCORES, BY GRADE

Table D1. Scheffé multiple range results for gains in criterion-referenced, K–12 math level of mastery scores, by grade, from data treated with the listwise deletion missing data technique

Mean	Grade	7	8	5	4	9	10	11	6	12	3	K	1	2
9.05	7	–												
10.51	8		–											
12.22	5			–										
12.53	4				–									
13.33	9					–								
14.44	10						–							
15.70	11							–						
16.41	6								–					
19.08	12									–				
21.08	3	*									–			
23.57	K	*										–		
33.19	1	*	*	*	*	*	*	*	*				–	
39.95	2	*	*	*	*	*	*	*	*	*	*	*		–

*$p < .05$.

Table D2. Scheffé multiple range results for gains in criterion-referenced, K–12 math level of mastery scores, by grade, from data treated with the grand-mean substitution missing data technique

	Grade	7	8	5	4	9	10	6	11	12	3	K	1	2
Mean	Grade													
9.49	7	–												
10.21	8		–											
10.29	5			–										
12.82	4				–									
14.10	9					–								
15.11	10						–							
15.85	6							–						
17.13	11								–					
18.79	12									–				
21.11	3	*	*	*							–			
22.60	K	*	*	*								–		
34.69	1	*	*	*	*	*	*	*	*	*	*	*	–	
38.88	2	*	*	*	*	*	*	*	*	*	*	*	*	–

*$p < .05$.

Table D3. Scheffé multiple range results for gains in criterion-referenced, K–12 math level of mastery scores, by grade, from data treated with the cell mean substitution missing data technique

Mean	Grade	7	5	8	4	9	10	6	11	12	3	K	1	2
9.42	7	–												
9.71	5		–											
9.91	8			–										
12.28	4				–									
12.81	9					–								
14.36	10						–							
16.14	6							–						
16.52	11								–					
16.85	12									–				
20.76	3	*	*	*							–			
22.94	K	*	*	*	*							–		
34.81	1	*	*	*	*	*	*	*	*	*	*	*	–	
39.86	2	*	*	*	*	*	*	*	*	*	*	*	*	–

*$p < .05$.

Table D4. Scheffé multiple range results for gains in criterion-referenced, K–12 math level of mastery scores, by grade, from data treated with the simple regression missing data technique

	Grade	7	5	8	10	9	4	11	12	6	3	K	1	2
Mean	Grade													
9.43	7	–												
10.23	5		–											
10.63	8			–										
12.32	10				–									
12.92	9					–								
13.28	4						–							
14.18	11							–						
14.71	12								–					
16.21	6									–				
20.70	3	*	*	*							–			
22.98	K	*	*	*	*							–		
34.00	1	*	*	*	*	*	*	*	*	*	*		–	
38.08	2	*	*	*	*	*	*	*	*	*	*	*		–

*$p < .05$.

APPENDIX E. SCHEFFÉ MULTIPLE RANGE RESULTS FOR K-12 MATH POSTTEST LEVEL OF MASTERY SCORES, BY GRADE

Table E1. Scheffé multiple range results for criterion-referenced, K–12 math posttest level of mastery scores, by grade, from data treated with the listwise deletion missing data technique

	Grade	7	8	4	11	12	6	10	9	5	K	3	1	2
Mean	Grade													
53.08	7	–												
55.01	8		–											
57.43	4			–										
57.58	11				–									
58.77	12					–								
59.44	6						–							
62.65	10							–						
62.98	9								–					
64.75	5									–				
69.61	K	*	*				*				–			
76.59	3	*	*	*	*		*	*	*			–		
83.82	1	*	*	*	*	*	*	*	*	*			–	
89.15	2	*	*	*	*	*	*	*	*	*	*			–

*$p < .05$.

Table E2. Scheffé multiple range results for criterion-referenced, K-12 math posttest level of mastery scores, by grade, from data treated with the pairwise deletion missing data technique

Mean	Grade	7	8	12	4	11	6	5	9	10	K	3	1	2
53.42	7	–												
54.46	8		–											
55.43	12			–										
56.99	4				–									
57.33	11					–								
58.66	6						–							
61.05	5							–						
62.38	9								–					
62.61	10									–				
68.91	K	*	*								–			
76.13	3	*	*		*	*	*	*	*	*		–		
85.08	1	*	*	*	*	*	*	*	*	*	*		–	
89.19	2	*	*	*	*	*	*	*	*	*	*	*		–

*$p < .05$.

Table E3. Scheffé multiple range results for criterion-referenced, K–12 math posttest level of mastery scores, by grade, from data treated with the grand-mean substitution missing data technique

	Grade	7	8	4	6	11	5	9	12	10	K	3	1	2
Mean	Grade													
53.57	7	–												
54.94	8		–											
57.65	4			–										
58.79	6				–									
60.27	11					–								
61.25	5						–							
62.89	9							–						
62.90	12								–					
63.12	10									–				
68.68	K	*	*								–			
75.96	3	*	*	*	*	*	*	*	*	*		–		
84.72	1	*	*	*	*	*	*	*	*	*	*		–	
87.86	2	*	*	*	*	*	*	*	*	*	*	*		–

*$p < .05$.

Table E4. Scheffé multiple range results for criterion-referenced, K–12 math posttest level of mastery scores, by grade, from data treated with the cell-mean substitution missing data technique

	Grade	7	8	12	4	11	6	5	9	10	K	3	1	2
Mean	Grade													
53.42	7	–												
54.46	8		–											
55.43	12			–										
56.99	4				–									
57.33	11					–								
58.66	6						–							
61.05	5							–						
62.38	9								–					
62.61	10									–				
68.91	K	*	*	*	*	*					–			
76.13	3	*	*	*	*	*	*	*	*	*		–		
85.08	1	*	*	*	*	*	*	*	*	*	*		–	
89.19	2	*	*	*	*	*	*	*	*	*	*	*		–

*$p < .05$.

Table E5. Scheffé multiple range results for criterion-referenced, K–12 math posttest level of mastery scores, by grade, from data treated with the simple regression missing data technique

Mean	Grade	7	8	11	12	4	6	10	5	9	K	3	1	2
53.52	7	–												
54.91	8		–											
55.13	11			–										
55.78	12				–									
57.82	4					–								
58.62	6						–							
59.92	10							–						
60.90	5								–					
61.16	9									–				
69.39	K	*	*	*	*						–			
75.98	3	*	*	*	*	*	*	*	*	*		–		
84.81	1	*	*	*	*	*	*	*	*	*	*		–	
88.34	2	*	*	*	*	*	*	*	*	*	*	*		–

*$p < .05$.

APPENDIX F. SCHEFFÉ MULTIPLE RANGE RESULTS FOR GAINS IN K-12 LANGUAGE ARTS LEVEL OF MASTERY SCORES, BY GRADE

Table F1. Scheffé multiple range results for gains in criterion-referenced, K–12 language arts level of mastery scores, by grade, from data treated with the listwise deletion missing data technique

	Grade	10	11	7	5	9	4	6	8	12	2	3	1	K
Mean	Grade													
-1.31	10	–												
2.54	11		–											
5.25	7			–										
5.74	5				–									
6.35	9					–								
6.82	4						–							
7.16	6							–						
8.35	8								–					
10.00	12									–				
15.08	2	*	*	*							–			
17.24	3	*	*	*	*	*	*	*				–		
32.55	1	*	*	*	*	*	*	*	*	*	*	*	–	
45.45	K	*	*	*	*	*	*	*	*	*	*	*		–

*$p < .05$.

Table F2. Scheffé multiple range results for gains in criterion-referenced, K–12 language arts level of mastery scores, by grade, from data treated with the grand-mean substitution missing data technique

	Grade	10	9	7	5	4	6	11	8	12	2	3	1	K
Mean	Grade													
-1.47	10	–												
4.73	9		–											
5.31	7			–										
6.10	5				–									
6.87	4					–								
6.95	6						–							
8.28	11							–						
8.43	8	*							–					
9.96	12	*								–				
16.73	2	*	*	*	*						–			
17.41	3	*	*	*	*	*	*					–		
31.83	1	*	*	*	*	*	*	*	*	*	*	*	–	
41.39	K	*	*	*	*	*	*	*	*	*	*	*		–

*$p < .05$.

Table F3. Scheffé multiple range results for gains in criterion-referenced, K-12 language arts level of mastery scores, by grade, from data treated with the cell-mean substitution missing data technique

	Grade	10	11	5	9	7	6	4	8	12	2	3	1	K
Mean	Grade													
-1.88	10	–												
2.70	11		–											
4.13	5			–										
4.47	9				–									
5.24	7					–								
6.49	6						–							
6.94	4							–						
8.27	8	*							–					
10.91	12	*								–				
15.33	2	*	*	*	*	*					–			
16.85	3	*	*	*	*	*	*	*				–		
36.18	1	*	*	*	*	*	*	*	*	*	*	*	–	
44.25	K	*	*	*	*	*	*	*	*	*	*	*		–

*$p < .05$.

Table F4. Scheffé multiple range results for gains in criterion-referenced, K–12 language arts level of mastery scores, by grade, from data treated with the simple regression missing data technique

	Grade	10	9	7	11	5	4	6	8	12	2	3	1	K
Mean	Grade													
-2.15	10	–												
5.14	9		–											
5.25	7			–										
5.88	11				–									
6.26	5					–								
7.36	4						–							
7.45	6							–						
8.49	8	*							–					
10.21	12	*								–				
15.40	2	*	*	*	*						–			
17.09	3	*	*	*	*	*	*					–		
26.39	1	*	*	*	*	*	*	*	*	*	*		–	
39.99	K	*	*	*	*	*	*	*	*	*	*	*	*	–

*$p < .05$.

APPENDIX G. SCHEFFÉ MULTIPLE RANGE RESULTS FOR K-12 LANGUAGE ARTS POSTTEST LEVEL OF MASTERY SCORES, BY GRADE

Table G1. Scheffé multiple range results for criterion-referenced, K–12 language arts posttest level of mastery scores, by grade, from data treated with the listwise deletion missing data technique

Mean	Grade	11	10	12	9	7	8	6	4	1	K	5	3	2
34.07	11	–												
40.18	10		–											
45.38	12			–										
54.51	9	*	*		–									
61.29	7	*	*			–								
62.37	8	*	*				–							
63.95	6	*	*					–						
70.33	4	*	*	*	*				–					
71.22	1	*	*	*	*					–				
72.56	K	*	*	*	*						–			
74.51	5	*	*	*	*	*	*					–		
82.76	3	*	*	*	*	*	*	*	*				–	
92.02	2	*	*	*	*	*	*	*	*	*	*	*		–

*$p < .05$.

Table G2. Scheffé multiple range results for criterion-referenced, K–12 language arts posttest level of mastery scores, by grade, from data treated with the pairwise deletion missing data technique

	Grade	11	10	12	9	7	8	6	4	1	5	K	3	2
Mean	Grade													
35.50	11	–												
39.24	10		–											
48.41	12	*		–										
52.35	9	*	*		–									
61.30	7	*	*	*		–								
62.35	8	*	*	*			–							
62.94	6	*	*	*	*			–						
69.84	4	*	*	*	*				–					
71.43	1	*	*	*	*					–				
71.75	5	*	*	*	*						–			
72.99	K	*	*	*	*							–		
82.38	3	*	*	*	*	*	*	*	*				–	
91.74	2	*	*	*	*	*	*	*	*	*	*	*		–

*$p < .05$.

Table G3. Scheffé multiple range results for criterion-referenced, K-12 language arts posttest level of mastery scores, by grade, from data treated with the grand-mean substitution missing data technique

Mean	Grade	10	11	12	9	7	8	6	1	4	5	K	3	2
40.83	10	–												
42.23	11		–											
49.17	12			–										
52.81	9	*	*		–									
61.32	7	*	*	*		–								
62.36	8	*	*	*			–							
62.91	6	*	*	*	*			–						
69.06	1	*	*	*	*				–					
69.20	4	*	*	*	*					–				
71.02	5	*	*	*	*						–			
72.61	K	*	*	*	*							–		
82.07	3	*	*	*	*	*	*	*	*	*			–	
89.76	2	*	*	*	*	*	*	*	*	*	*	*		–

*$p < .05$.

Table G4. Scheffé multiple range results for criterion-referenced, K–12 language arts posttest level of mastery scores, by grade, from data treated with the cell-mean substitution missing data technique

	Grade	11	10	12	9	7	8	6	4	1	5	K	3	2
Mean	Grade													
35.50	11	–												
39.24	10		–											
48.41	12	*		–										
52.35	9	*	*		–									
61.30	7	*	*	*		–								
62.35	8	*	*	*	*		–							
62.94	6	*	*	*	*			–						
69.84	4	*	*	*	*				–					
71.43	1	*	*	*	*					–				
71.75	5	*	*	*	*	*					–			
72.99	K	*	*	*	*	*						–		
82.38	3	*	*	*	*	*	*	*	*				–	
91.74	2	*	*	*	*	*	*	*	*	*	*	*		–

*$p < .05$.

Table G5. Scheffé multiple range results for criterion-referenced, K–12 language arts posttest level of mastery scores, by grade, from data treated with the simple regression missing data technique

Mean	Grade Grade	11	10	12	9	7	8	6	1	4	5	K	3	2
39.03	11	–												
40.33	10		–											
48.97	12			–										
52.81	9	*	*		–									
61.20	7	*	*	*		–								
62.44	8	*	*	*			–							
63.00	6	*	*	*	*			–						
65.13	1	*	*	*	*				–					
70.26	4	*	*	*	*					–				
71.19	5	*	*	*	*						–			
72.64	K	*	*	*	*							–		
82.35	3	*	*	*	*	*	*	*	*	*			–	
91.02	2	*	*	*	*	*	*	*	*	*	*	*	*	–

*p<.05.

APPENDIX H. MEAN DEVIATION AND MEAN ABSOLUTE DEVIATION SCORES FOR MEANS, STANDARD DEVIATIONS, AND CORRELATIONS

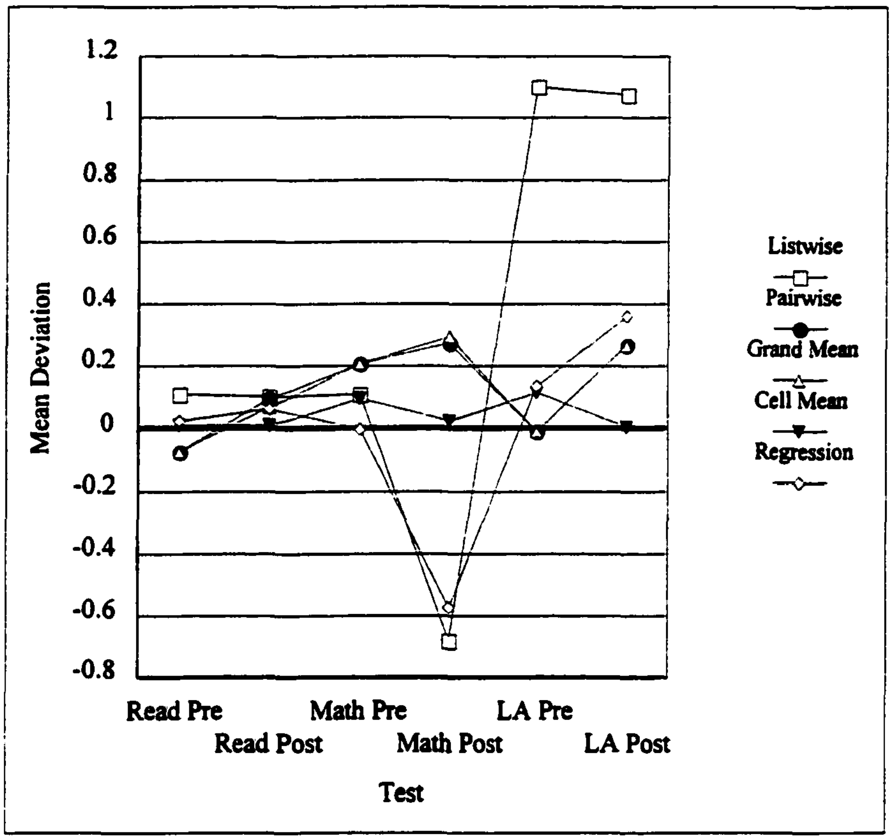

Figure H1. Mean deviation scores for test means

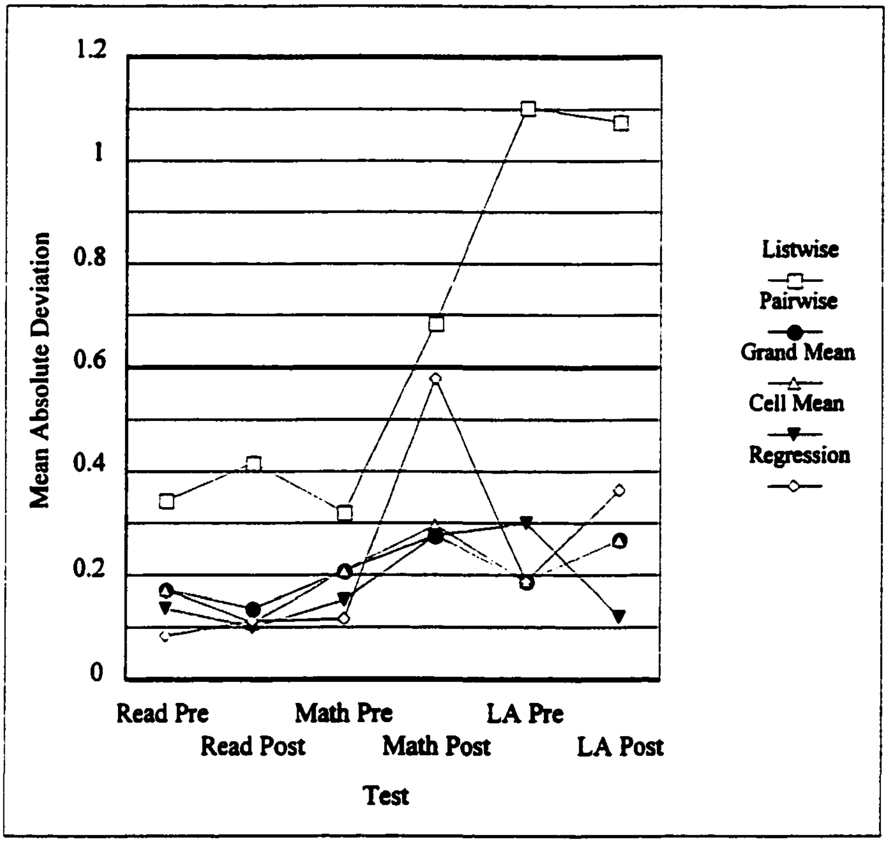

Figure H2. Mean absolute deviation scores for test means

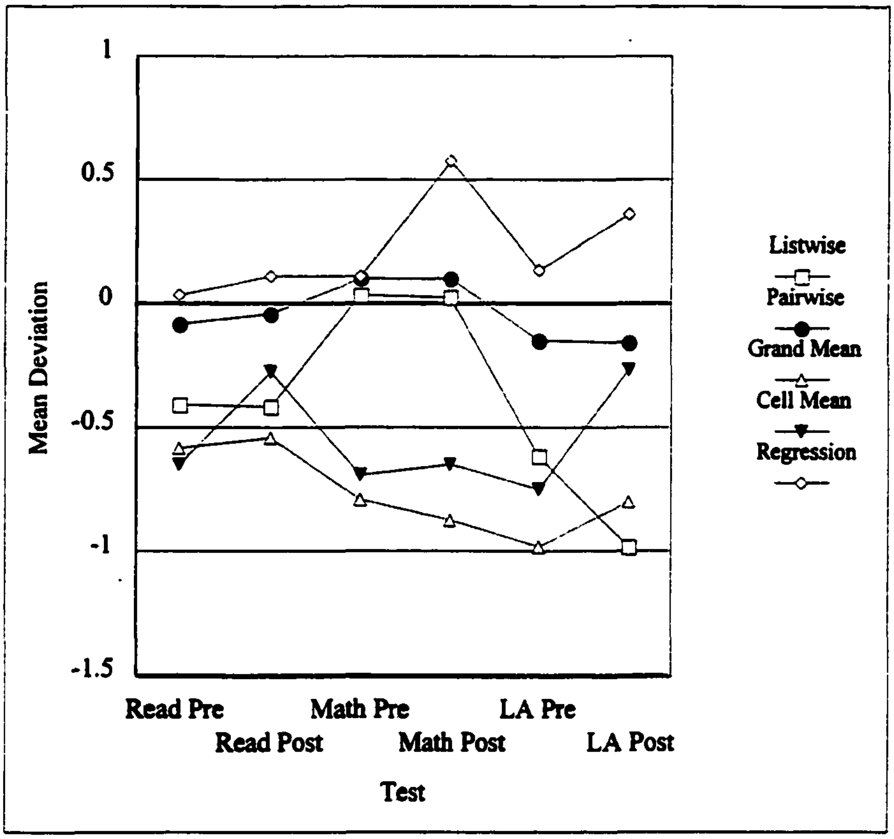

Figure H3. Mean deviation scores for standard deviations

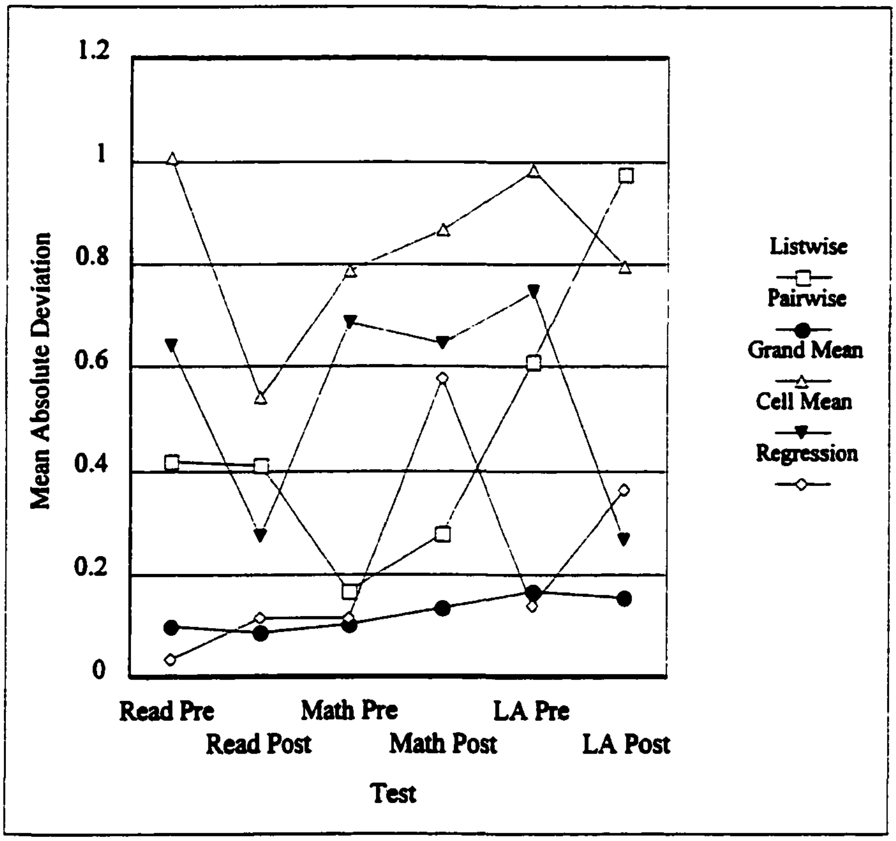

Figure H4. Mean absolute deviation scores for standard deviations

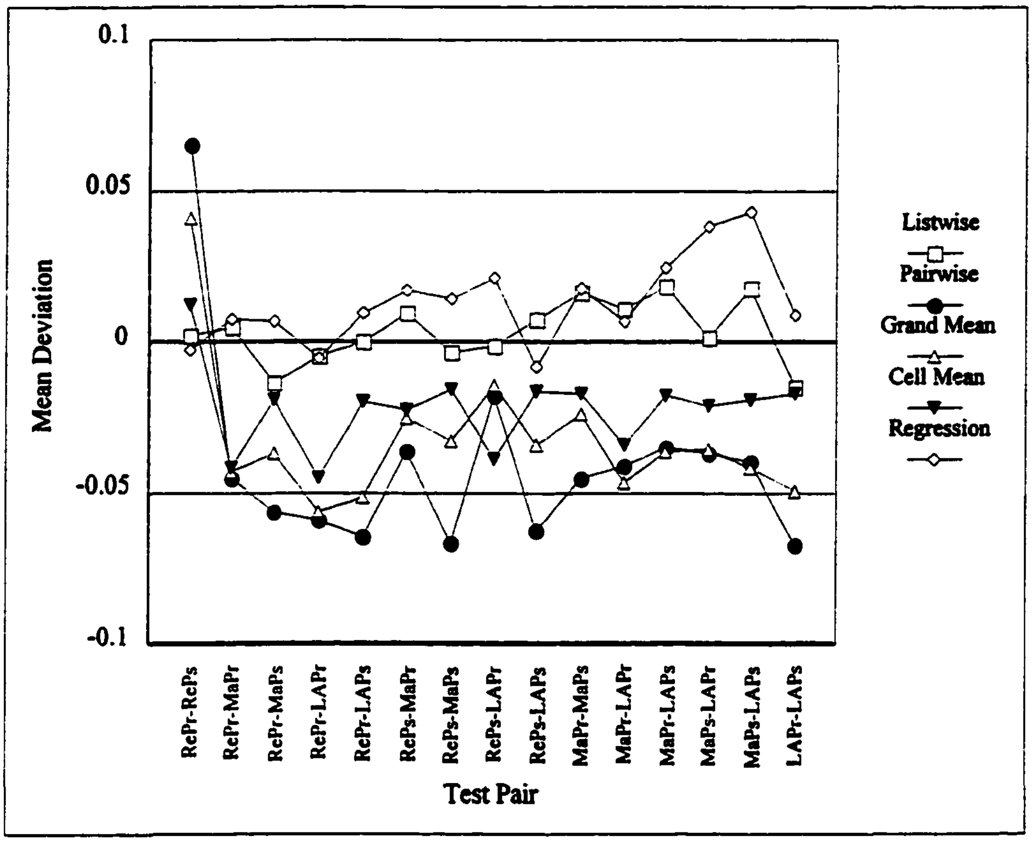

Figure H5. Mean deviation measures for correlations

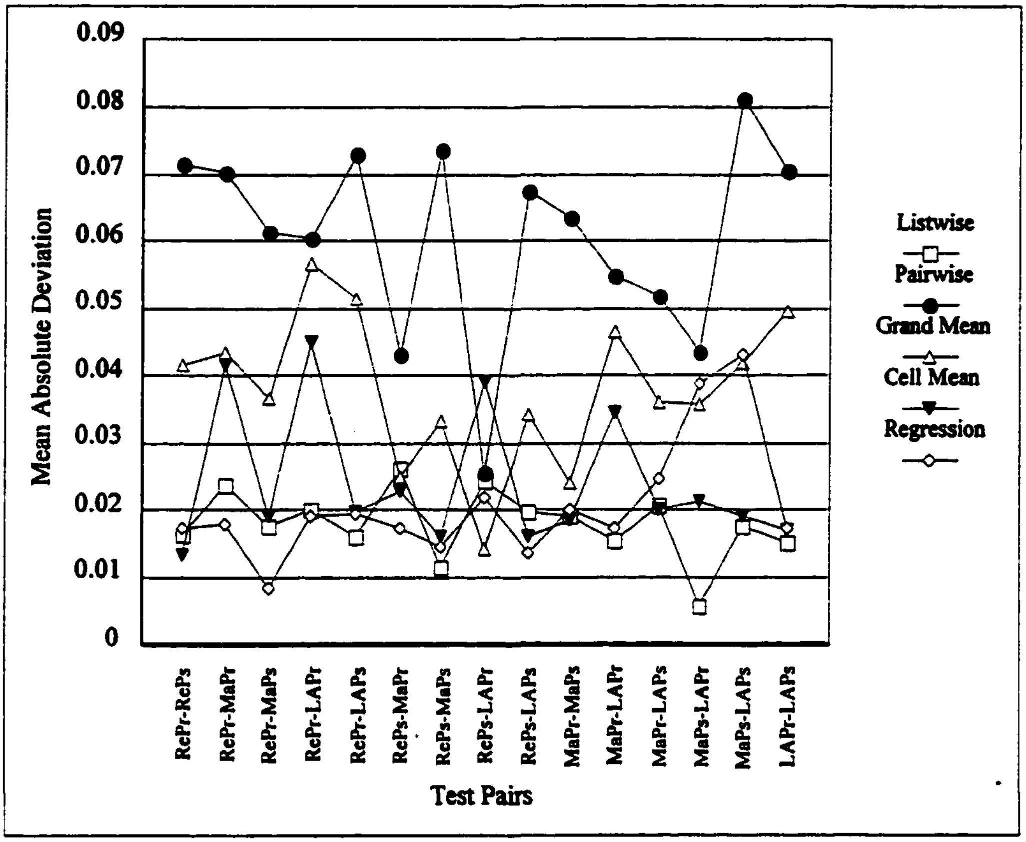

Figure H6. Mean absolute deviation measures for correlations

BIBLIOGRAPHY

American Federation of Teachers. (1999). *Making standards matter 1998: Executive summary* [On-line]. Available: http://www.aft.org/edissues/standards98/index.htm.

Armoogum, J., & Madre, J. (1998). Weighting or imputations? The example of nonresponses for daily trips in the French NPTS. *Journal of Transportation and Statistics, 1*(3), 53–63.

Associated Press. (1999). *Supreme court rejects sampling* [On-line]. ABCNEWS.com. Available: http://www.abcnes.go.com/sections/us/DailyNews/census990124.html.

Ayers, W. (1993). *To teach: The journey of a teacher.* New York: Teachers College Press.

Baker, E. L., & Linn, R. L. (1995). United States. In Organisation for Economic Co-operation and Development, *Performance standards in education: In search of quality.* (pp. 197–210). Paris, France: Author.

Basilevsky, A., Sabourin, D., Hum, D., & Anderson, A. (1985). Missing data estimators in the general linear model: An evaluation of simulated data as an experimental design. *Communications in Statistics: Simulation & Computation, 14*(1), 371–394.

Beaton, A. E. (1997). Missing scores in survey research. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: An international handbook* (2nd ed.). (pp. 763–766). Cambridge, UK: Cambridge University Press.

Berube, M. B. (1994). *American school reform: Progressive, equity, and excellence movements, 1883-1993.* Westport, CT: Greenwood Press.

Biskupic, J. (1998). High court to rule on census sampling. *Washington Post* [On-line]. Available: wysiwig://68/http://wpni.com/wp-srv/politics/daily/sept98/census11.htm.

Borman, K. M., & Greenman, N. P. (Eds.) (1994). *Changing American education: Recapturing the past or inventing the future?* Albany, NY: State University of New York Press.

Boyd, W. L. (1990). Balancing control and autonomy in school reform: The politics of perestroika. In J. Murphy (Ed.), *The Educational Reform Movement of the 1980s: Perspectives and Cases.* (pp. 85–96). Berkeley, CA: McCutchan Publishing Corporation.

Center for Policy Research in Education. (1989). *State education reform in the 1980s. CPRE Policy Briefs.* (Report No. CPRE-RB-03-11/89). Washington, DC: Office of Educational Research and Improvement. (ERIC Document Reproduction Service No. ED 342 105)

Clotfelter, C. T., & Ladd, H. F. (1996). Recognizing and rewarding success in public schools. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education.* (pp. 23–64). Washington, DC: The Brookings Institution.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

DiConti, V. D. (1996). *Interest groups and education reform: The latest crusade to restructure the schools*. Lanham, MD: University Press of America, Inc.

Education Week. (1999). Executive summary: Demanding results. In *Quality Counts '99* [On-line]. Available: http://www.edweek.org/sreports/qc99/exsum.htm.

Elmore, R. F., Abelman, C. H., & Huhrman, S. H. (1996). The new accountability in state education reform: From process to performance. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education*. (pp. 65–89). Washington, DC: The Brookings Institution.

Feir, R. E. (April 1995). *Political and social roots of education reform: A look at the state in the mid-1980s*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA: American Educational Research Association. (ERIC Document Reproduction Service No. ED 382 925)

Finn, C. E., Jr. (1995). Who's afraid of the big bad test. In D. Ravitch (Ed.), *Debating the future of American education: Do we need national standards and assessments?* (pp. 120–144). Washington, DC: The Brookings Institution.

Finn, C. E., Jr., & Rebarber, T. (Eds.). (1992). *Education reform in the '90s*. New York: Macmillan Publishing Company.

Firestone, W. A. (1990). Continuity and incrementalism after all: State responses to the excellence movement. In J. Murphy (Ed.), *The educational reform movement of the 1980s: Perspectives and cases*. (pp. 143–166). Berkeley, CA: McCutchan Publishing Corporation.

Frymier, J. (1996). *Accountability in education: Still an evolving concept* (Fastback 395). Bloomington, IN: Phi Delta Kappa Educational Foundation.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction* (6th ed.). White Plains, NY: Longman Publishers.

Gleason, T. C., & Staelin, R. (1975). A proposal for handling missing data. *Psychometrika, 40*(2), 229–252.

Greenman, N. P. (1994). Not all caterpillars become butterflies: Reform and restructuring as educational change. In K. M. Borman & N. P. Greenman (Eds.), *Changing American education: Recapturing the past or inventing the future?* (pp. 3–32). Albany, NY: State University of New York Press.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society* (series B), *30*(1), 67–82.

Haney, W., & Madaus, G. (1989). Searching for alternatives to standardized tests: Whys, whats, and whithers. *Phi Delta Kappan, 70*(9), 683-687.

Harnisch, D. L., & Mabry, L. (1993). Issues in the development and evaluation of alternative assessments. *Journal of Curriculum Studies, 25*(2), 179-187.

Hegamin-Younger, C., & Forsyth, R. (1998). A comparison of four imputation procedures in a two-variable prediction system. *Educational and Psychological Measurement, 58*(2), 197-210.

Henry, N. W. (1995). *On applying statistical methods: Losing local control.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. 387 514)

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998). *Applied statistics for the behavioral sciences.* Boston, MA: Houghton Mifflin Company.

Jennings, J. F. (1998). *Why national standards and tests? Politics and the quest for better schools.* Thousand Oaks, CA: Sage Publications, Inc.

Kaiser, J., & Tracy, D. B. (1988). *Estimation of missing values by predicted scores.* Paper presented at the annual meeting of the American Statistical Association, New Orleans, LA. (ERIC Document Reproduction Service No. 298 146)

Kalton, G. (1983). *Compensating for missing survey data.* Research Report Series, Institute for Social Research, University of Michigan: Ann Arbor, MI.

Kim, J., & Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods and Research, 6*(2), 215-240.

Kirst, M. W. (1988). Recent state education reform in the United States: Looking backward and forward. *Educational Administration Quarterly, 24*(3), 319-328.

Kirst, M. W. (1990). *Accountability: Implications for state and local policymakers.* Washington, DC: U.S. Department of Education.

Kogan, M. (1986). *Education accountability: An analytic overview.* London: Hutchinson & Co., Ltd.

Kromrey, J. D., & Hines, C. V. (1991). *Randomly missing data in multiple regression: An empirical comparison of common missing data treatments.* Paper presented at the annual meeting of the Eastern Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. ED 329 593)

Lessinger, L. M. (1971). Accountability in perspective. In L. M. Lessinger & R. W. Tyler (Eds.), *Accountability in education.* (pp. 7–14). Worthington, OH: Charles A. Jones Publishing Company.

Leung, R. (1998). *Stakes high in case over statistical sampling* [On-line]. ABCNEWS.com. Available: http://www.abcnes.go.com/sections/us/DailyNews/census981130_final.html.

Linn, R. L. (1992). Achievement Testing. In M. C. Alkin (Ed.), *Encyclopedia of educational research.* (pp. 1–12). New York: American Educational Research Association.

Lipsky, D. K. (1992). We need a third wave of education reform. *Social Policy, 22*(3), 43–45.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: John Wiley & Sons Publications.

Little, R. J. A., & Rubin, D. B. (1989–1990). The analysis of social science data with missing values. *Sociological Methods and Research, 18*(2&3), 292–326.

Little, R. J. A., & Schenker, N. (1995). Missing Data. In G. Arminger, C. Clogg, & M. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39–75). New York: Plenum Press.

Macpherson, R. J. S. (Ed.). (1995). Accountability research in education: Current developments. *International Journal of Educational Research, 23*(6), 479–492.

Malhotra, N. K. (1987). Analyzing marketing research data with incomplete information on the dependent variable. *Journal of Marketing Research, 24*(1), 74–84.

Manatt, R. P. (1993). The changing paradigm of outcomes and assessment. *International Journal of Education Reform, 2*(1), 86–95.

Marzano, R. J., Pickering, D., & McTighe, J. (1993). *Assessing student outcomes: Performance assessment using the dimensions of learning model.* Alexandria, VA: Association for Supervision and Curriculum Development.

Maxwell, A. (1997). *Census chief defends sampling* [On-line]. GovExec.com. Available: http://govexec.com/dailyfed/0897/080697a1.htm.

Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review, 16*(3), 283–301.

Michaels, K. (1988). Caution: Second wave reform taking place. *Educational Leadership, 45*(5), 3.

Mitchell, K. J. (1996). *Reforming and conforming: NASDC Principals discuss school accountability systems.* Santa Monica, CA: Rand Corporation.

Moss, P. A., & Schutz, A. (1999). Risking frankness in educational assessment. *Phi Delta Kappan, 80*(9), 680–687.

Murphy, J. (Ed.) (1990). *The educational reform movement of the 1980s: Perspectives and cases.* Berkeley, CA: McCuthan Publishing Corporation.

Murphy, J. (1992). Restructuring America's schools: An overview. In C. E. Finn & T. Rebarber (Eds.), *Education reform in the '90s.* (pp. 3–22). New York: Macmillan Publishing Company.

National Education Association. (1997, June). Accountability and school reform. *National Issues in Education* [On-Line serial, No. 6]. Available: http://www.nea.org/helpfrom/achieve/accountability/school.html.

Newman, F. M., King, M. B., & Rigdon, M. (1997). Accountability and school performance: Implications from restructuring schools. *Harvard Educational Review, 67*(1), 41–74.

Olden, A., & Marsh, D. (1990). Local response to the 1980s state education reforms: New patterns of local and state interaction. In J. Murphy (Ed.), *The educational reform movement of the 1980s: Perspectives and cases.* (pp. 167–186). Berkeley, CA: McCutchan Publishing Corporation.

Olson, L. (1999). Shining a spotlight on results. In *Quality Counts '99* [On-line]. Available: http://www.edweek.org/sreports/qc99/ac/mc/mc-intro.htm.

O'Neil, J. (1992). Putting performance assessment to the test. *Educational Leadership, 49*(8), 14–19.

Patton, M. Q. (1982). *Practical evaluation.* Newbury Park, CA: Sage Publications.

Parker, F., & Parker, B. J. (1995). A historical perspective on school reform. *The Education Forum, 59*(3), 278–286.

Pipho, C. (1996). The standards parade. *Phi Delta Kappan, 77*(10), 655, 701.

Pipho, C. (1999). The search for better education information. *Phi Delta Kappan, 79*(7), 485–486.

Plank, D. N., & Ginsberg, R. (1990). Catch the wave: Reform commissions and school reform. In J. Murphy (Ed.), *The educational reform movement of the 1980s: Perspectives and cases.* (pp. 121–142). Berkeley, CA: McCutchan Publishing Corporation.

Plato, K. (1992). The politics of assessment reform: Implications for educators. *National Association of Secondary School Principals Bulletin, 76*(545), 41–49.

Porter, A. C. (1994). National standards and school improvement in the 1990s: Issues and promise. *American Journal of Education, 102*(4), 421–449.

Ravitch, D. (Ed.). (1995). *Debating the future of American education: Do we need national standards and assessments?* Washington, DC: The Brookings Institution.

Raymond, M. R. (1987). *An interactive approach to analyzing incomplete multivariate data.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 281 854)

Raymond, M. R., & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement, 47,* 13–26.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47,* 537–560.

Roth, P. L., Campion, J. E., & Jones, S. D. (1996). The impact of four missing data techniques on validity estimates in human resource management. *Journal of Business and Psychology, 11*(1), 101–112.

Schmoker, M. (1996). *Results: The key to continuous school improvement.* Alexandria, VA: Association for Supervision and Curriculum Development.

Sciara, F. J., & Jantz, R. K. (1972). *Accountability in American education.* Boston, MA: Allyn and Bacon, Inc.

Shanker, A. (1995). The case for high stakes and real consequences. In D. Ravitch (Ed.), *Debating the future of American education: Do we need national standards and assessments?* (pp. 145–153). Washington, DC: The Brookings Institution.

Shepard, L. A. (1992). Uses and abuses of testing. In M. C. Alkin (Ed.), *Encyclopedia of educational research.* (pp. 1477–1485). New York: American Educational Research Association.

Stake, R. (1999). The goods on American education. *Phi Delta Kappan, 80*(9), 668–679, 672.

Stiggins, R. J. (1994). *Student-centered classroom assessment.* New York: MacMillan College Publishing Company, Inc.

Thran, S. L., & Gillis, K. D. (1992). A comparison of imputation techniques in a physician survey. *American Statistical Association 1992 Proceedings of the Section on Survey Research Methods.* (pp. 211–220). Alexandria, VA: American Statistical Association.

Timm, N. H. (1970). The estimation of variance-covariance and correlation matrices from incomplete data. *Psychometrika, 35*(4), 417–437.

Tyler, R. W. (1971). Accountability in perspective. In L. M. Lessinger & R. W. Tyler (Eds.), *Accountability in education.* (pp. 1–6). Worthington, OH: Charles A. Jones Publishing Company.

Velotta, C. L. (1995). *Coping with missing data in educational research and evaluation.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 388 686)

Ward, T. J., Jr., & Clark H. T., III. (1991). A reexamination of public- versus private-school achievement: The case for missing data. *Journal of Educational Research, 84*(3), 153-163.

Warren, D. (1990). Passage of rites: On the history of educational reform in the United States. In J. Murphy (Ed.), *The educational reform movement of the 1980s: Perspectives and cases.* (pp. 57-82). Berkeley, CA: McCutchan Publishing Corporation.

Wiggins, G. (1989a). Teaching to the (authentic) test. *Educational Leadership, 46*(7), 41-47.

Wiggins. G. (1989b). Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 703-713.

Wilkerson, D. J. (1997). *The association of performance ratings of teachers and achievement of students in the classroom.* Doctoral dissertation, Iowa State University, Ames, IA.

Williamson, C. (1998). *Census 2000: A 1990 repeat? Concerns and implications for planners* [On-line]. American Planning Association. Available: http://www.planning.org/info/census.htm.

Witta, E. L. (1994). *Are values missing randomly in survey research?* Paper presented at the annual meeting of the Mid-South Educational Research Association, Nashville, TN. (ERIC Document Reproduction Service No. ED 389 727)

Witta, L., & Kaiser, J. (1991). *Four methods of handling missing data with the 1984 General Social Survey.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Lexington, KY. (ERIC Document Reproduction Service No. 339 755)

Worthen, B. R. (1993). Critical issues that will determine the future of alternative assessment. *Phi Delta Kappan, 74*(6), 444-456.

ACKNOWLEDGMENTS

Tasks like writing a dissertation are very rarely accomplished in isolation. I would like to thank Dr. Richard Manatt, chairman of my doctoral committee, for his patience and encouragement. His ongoing support and occasional "kicks in the seat of the pants" are responsible for the product you are now reading. Dr. Manatt's willingness to oversee a work he described as "esoteric" speaks well of his ability to work with students, his wealth of knowledge, and his dedication.

My appreciation also extends to the members of my doctoral committee, Dr. Donna Merkley, Dr. William Poston, Dr. Howard Shapiro, and Dr. Shirley Stow, for their patience, encouragement, and cooperation as I completed this work while working full time outside of academia.

I would like to extend a brief, yet heart-felt, thank you to Dr. Anton (Tony) Netusil. Though he was unable to serve on my committee through the completion of the dissertation, he helped me see that being analytically minded wasn't necessarily a bad thing.

I owe my deepest debt to my family, who put up with my not being around and never once complained. My mother and father instilled a thirst for knowledge for which I will always be grateful. I only hope I can transfer a similar drive to my children. To my wife I can only say, "God bless you." She stuck with me throughout the entire graduate school experience and was always there with words of encouragement and support.